UNIVERSIDAD DE LA REPÚBLICA

Facultad de Ciencias Económicas y de Administración

Licenciatura en Estadística

Informe de Pasantía

Una revisión de los modelos de conteo con excesos de ceros.

Eloísa Martínez Calcaterra

Pamela Vaucher Silva

Tutores:

Ramón Alvarez

Ana Coimbra

Montevideo, Diciembre 2017.

UNIVERSIDAD DE LA REPÚBLICA

FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

El tribunal docente integrado por los abajo firmantes aprueba el trabajo de Pasantía:

Una revisión de los modelos de conteo con excesos de ceros.

Pamela Vaucher Silva - Eloísa Martínez Calcaterra

Tutores:

Ramón Alvarez

Ana Coímbra

Licenciatura en Estadística

Puntaje .....

Tribunal

Profesor.....(nombre y firma).

Profesor.....(nombre y firma).

Profesor.....(nombre y firma).

Profesor.....(nombre y firma).

Fecha.....

Resumen

El objetivo de este trabajo es encontrar modelos predictivos que describan el conteo de C, P, O y CPO que son indicadores de patología bucal. El indicador CPO señala la experiencia de caries tanto presentes como pasadas, ya que es la suma del número de dientes cariados (C), número de dientes perdidos por la enfermedad (P) y número de dientes obturados(O) para cada individuo.

Los datos con los que se trabaja provienen del primer relevamiento en salud oral llevado a cabo por el Servicio de Epidemiología y Estadística de Facultad de Odontología, coordinado conjuntamente con docentes del Instituto de Estadística de Facultad de Ciencias Económicas. Es un estudio realizado en el período 2010-2011 con un diseño de muestreo probabilístico complejo (el cual no será considerado en este trabajo) a la población joven y adulta urbana en sus domicilios, tanto en Montevideo como en el Interior del país. Se relevó información de variables sociodemográficas así como variables clínicas.

Los datos de conteo muestran, además de sobredispersión, una gran cantidad de ceros, por lo que se trabaja con Modelos Lineales Generalizados con excesos de ceros. Estos son modelos de conteo mixtos ya que combinan variables truncadas.

Se modelan los componentes del índice CPO por separado así como el propio CPO, llegando a verificar que presentan distintos comportamientos en cuanto a su distribución y a las variables explicativas que inciden en su conteo.

Palabras clave: CPO, exceso de ceros, modelos de conteo, modelos lineales generalizados, sobredispersión.

IV

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Es de interés para los profesionales de la odontología, medir el grado de salud bucal relacionada con la enfermedad de caries dental de los individuos y considerando un grupo de variables de tipo sociales, económicas, demográficas y clínicas estudiar si existe relación entre ellas y la enfermedad caries.

El índice más utilizado para medir la prevalencia de dicha enfermedad es el Índice CPO, que debe su nombre a las primeras letras de Cariado, Perdido y Obturado. El mismo fue propuesto por Klein, Palmer y Knutson (9) en el año 1935, durante un estudio del estado dental y la necesidad de tratamiento de niños asistentes a escuelas primarias en Hagerstown, Maryland, EUA. Este índice es un indicador que cuantifica la experiencia de la enfermedad Caries Dental tanto presente como pasada, debido a que toma en cuenta los dientes con la enfermedad presente, los que tuvieron un tratamiento previo e incluso los que han sido extraídos a causa de la misma.

De éste modo, el índice CPO del individuo j se obtiene sumando la cantidad de dientes permanentes Cariados (C), Perdidos (P) y Obturados (O) de cada individuo.

$$CPO_j = \sum_{i=1}^{32} C_i + \sum_{i=1}^{32} P_i + \sum_{i=1}^{32} O_i \tag{1.1}$$

donde

- **C**: Cariado - La enfermedad está presente y la lesión activa
- **P**: Perdido - La pieza fue perdida por caries dental
- **O**: Obturado - La pieza recibió tratamiento y la enfermedad ha sido curada.

De esta forma, C_i vale 1 si la pieza i presenta caries y cero si no, P_i vale 1 si ha sido perdida por la enfermedad y cero si no, y O_i vale 1 si ha sido curada y cero si no, de modo que el índice CPO puede tomar valores de 0 a 32, ya que se contabilizan 32 piezas dentales en el caso de que se tengan los terceros molares, a los que se llama "muelas de juicio".

A partir de éste índice y realizando algunas modificaciones al mismo se propusieron luego una variedad de indicadores con el mismo propósito. En 1944 Gruebbel (5) propone el CPO-d que se obtiene de igual manera que el CPO pero toma en cuenta sólo los dientes temporales, por lo que se consideran 20 piezas e individuos menores de 12 años. Se puede encontrar otras variaciones en textos de Odontología y de Salud Bucal.

Debido a la variabilidad oculta que presenta el indicador, ya que un valor de CPO=10 puede referirse tanto a 10 piezas perdidas como a 5 cariadas y 5 obturadas, en el presente trabajo se analiza cada componente individualmente.

Los datos utilizados para el estudio son los correspondientes al primer relevamiento epidemiológico llevado a cabo en Uruguay durante los años 2010-2011(12) por parte de la Facultad de Odontología de la Universidad de la República, auspiciado por el Ministerio de Salud Pública y coordinado conjuntamente con docentes del Instituto de Estadística de Facultad de Ciencias Económicas y de Administración, y basado en la metodología propuesta por la Organización Mundial de la Salud (OMS)[1].

[1] http://www.who.int/about/es/

La muestra consta de 1485 individuos relevados, de los cuales 922 pertenecen a Montevideo y el resto a 14 ciudades del interior que tienen más de 20.000 habitantes. Los tramos etarios en los que se divide la población de estudio son de 15 a 24, 35 a 44 y 65 a 74.

Objetivos

El objetivo general es encontrar una forma de explicar las variables C, P, y O así como también del indicador CPO a partir de un conjunto de variables socioeconómicas que se consideran importantes en el resultado de los mismos y así encontrar un modelo adecuado para estudiar el comportamiento de C, P, O y CPO.

Como objetivos específicos se plantea:

- Encontrar las distribuciones que mejor se adapten a las variables C, P, O y CPO.
- Ver si las variables económicas, sociales y demográficas que son significativas para explicar una variable son las mismas o no para las otras variables odontológicas.

Estructura del Trabajo

El presente trabajo consta de 5 capítulos. En el primer capítulo se realiza una introducción a lo que será el mismo, así como también una breve explicación de lo que es y cómo surge el índice CPO y una introducción de los datos empleados. En el segundo capítulo, se presenta la metodología estadística utilizada. Al inicio del mismo se muestran los procedimientos básicos para el análisis de regresión, mostrándose métodos más complejos sobre el final. Se hace especial énfasis a los modelos de conteo con exceso de ceros y al final del capítulo se muestran los procedimientos para evaluar el ajuste del modelo de regresión. En el tercer capítulo se presentan los datos de la aplicación y se realiza una descripción de los mismos. También se realiza un

análisis de las variables que serán utilizadas para explicar el comportamiento de las variables C, P y O. En el capítulo 4 se muestran los resultados de la aplicación, y en el capítulo 5 se exponen las principales conclusiones y los pasos a seguir en futuros trabajos.

Capítulo 2

Metodología

En este capítulo se describen los aspectos metodológicos estadísticos necesarios para el análisis de las variables C, P, O y CPO. En primer lugar se exponen los pasos a seguir para determinar las familias de distribuciones más apropiadas para representar la variable a explicar. Luego, se hace un recorrido por los diferentes tipos de Modelos pasando por Modelos de Regresión, Modelos Lineales Generalizados, y finalmente Modelos de Regresión con excesos de ceros. Sobre el final del capítulo se expone la metodología empleada para evaluar el modelo de regresión construído y la calidad del ajuste del mismo.

2.1. Determinación de la distribución de los datos

Para encontrar el modelo de conteo apropiado a fin de explicar el comportamiento de la variable C, P, O o CPO, es necesario determinar la distribución de la misma. Los pasos a seguir para estimar la distribución son (20):

1. Elegir las posibles familias de distribuciones de probabilidad que mejor ajuste a los datos.

2. Estimar los parámetros de la distribución de probabilidad seleccionada.

3. Evaluar la calidad del ajuste de la distribución.

2.1.1. Elección de las posibles familias de distribuciones que ajusten a los datos bajo estudio

Una forma de elegir las posibles familias de distribuciones que se supone que mejor representan a los datos, es por medio de análisis exploratorio de los datos a través de medidas de resumen univariadas o por medio de gráficos. El histograma, por ejemplo, permite comparar gráficamente las funciones de densidad teóricas con las empíricas. Pero esto puede ser muy subjetivo, por lo que se deben buscar métodos analíticos que sean más objetivos.

2.1.2. Estimación de parámetros de la función de distribución

La estimación de los parámetros de la función de distribución, $\theta \in \Theta$ asociados a una distribución de probabilidad se puede hacer por distintos métodos: Método de los Momentos (MM) o Máxima Verosimilitud (MV).

- Método de Momentos: Se igualan momentos poblacionales con momentos muestrales para hallar los parámetros. Sea Y una variable aleatoria con función de densidad o cuantía $f_y(y)$, el momento muestral de orden t para $t \in \mathbb{N}$ es:

$$E(Y^t) = \begin{cases} \sum_y y^t f_y(y;\theta) & \text{en el caso discreto} \\ \int_y y^t f_y(y;\theta)dy & \text{en el caso continuo} \end{cases}$$

- Método Máxima Verosimilitud: La función de verosimilitud es una función de los parámetros donde las y_i son dadas. Un estimador MV de un parámetro es aquel valor que maximiza la probabilidad de observar una determinada muestra.

  Función de Verosimilitud: $L(\theta) = \prod_{i=1}^{n} f(y_i; \theta)$

  Maximizar el logaritmo de esta función equivale a maximizar la función, lo cual resulta más simple. Se deriva respecto a θ el logaritmo de la verosimilitud y se iguala a cero para encontrar el máximo de la función. De esa ecuación se despejan los valores que serán las estimaciones de los parámetros.

2.1.3. Calidad del ajuste

Para estudiar el ajuste de una distribución de probabilidad a los datos se estudian los errores que resultan de aplicar la distribución de probabilidad elegida a la muestra, y además se emplean una serie de tests, conocidos como Tests de Bondad de Ajuste.

Las medidas de Bondad de Ajuste describen el ajuste de un conjunto de observaciones a una distribución de probabilidad. Se usan para comparar frecuencias empíricas con frecuencias teóricas; es decir, resumir la discrepancia entre los valores observados y los valores esperados. Existen medidas absolutas y relativas. Las medidas absolutas son las que consideran las diferencias entre el valor observado y el valor estimado y las medidas relativas son el cociente entre el error absoluto y el valor observado. Como medida absoluta se presenta por ejemplo:

$$\xi = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{2.1}$$

donde y_i es la frecuencia empírica y $\hat{y}_i$ es el valor ajustado. Y como medida relativa:

$$\delta = \frac{\xi}{\sum_{i=1}^{n} y_i/n} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} y_i} \tag{2.2}$$

Estas medidas muestran la diferencia que existe entre el valor observado y el valor esperado que resulta al aplicar la distribución de probabilidad que se desea probar. De este modo, cuanto más chica esta medida, más acertada la elección de la distribución. Se elige la función de distribución de probabilidades que presente el menor valor de medida absoluta o relativa.

Es posible evaluar la bondad de ajuste gráficamente representando la densidad teórica y el histograma juntos: cuanto más se asemeja el histograma de los datos observados al gráfico de la función de densidad o función de cuantía, mejor el ajuste.

Los tests de bondad de ajuste son una herramienta utilizada para probar si los datos que se estudian provienen de una distribución de probabilidad dada. Se realizan mediante pruebas de hipótesis de la forma:

H_0) La muestra proviene de la distribución indicada.

H_1) La muestra no proviene de dicha distribución.

Estas pruebas no dependen de la función de distribución.

La prueba Chi-cuadrado es un test de bondad de ajuste basado en una comparación de los valores observados y los valores esperados bajo H_0 cierta. Se trabaja con datos agrupados y se considera el ajuste de la frecuencia observada con la frecuencia esperada según la función de distribución de la hipótesis en cada grupo. Esta prueba se puede usar con cualquier función de distribución, tanto discreta como continua. Una desventaja de este test es que dado que se aplica con los datos agrupados el resultado de la prueba dependerá de cómo se agrupen los mismos. Otra desventaja es que para su implementación es necesario un tamaño de muestra suficientemente grande (por ejemplo > 50).

Para la prueba se dividen los datos en l grupos y el estadístico de prueba es el siguiente:

$$\chi^2 = \sum_{i=1}^{l} \frac{(O_i - E_i)^2}{E_i} \tag{2.3}$$

donde O_i es la frecuencia observada del grupo i y E_i el número esperado para el grupo i calculado por la función de distribución propuesta. Este estadístico tiene distribución χ^2 con $l-k-1$ grados de libertad, donde k es el número de parámetros estimados.

2.2. Análisis de regresión

El análisis de regresión es un proceso utilizado para conocer el efecto que una o varias variables independientes o predictoras causan sobre una variable dependiente o variable de respuesta. De esta manera, es de interés en el análisis de regresión explorar el cambio en el valor esperado de la variable dependiente Y cuando el valor de una de las variables predictoras X varía manteniendo las otras constantes. Se estima la función de regresión, que es la función del conjunto de variables independientes.

La relación entre la variable dependiente y la función de regresión puede ser lineal, dando paso a los Modelos de Regresión Lineal (MRL), o no lineales, dónde la técnica usada es Modelos Lineales Generalizados (MLG).

El modelo de regresión tiene la forma

$$E(Y/X) = f(x) + \epsilon \tag{2.4}$$

Donde ϵ es la diferencia entre el valor ajustado y el valor real y se conoce como "error

aleatorio". Cuando Y depende de una única variable regresora X y la relación es lineal, es clasificada como Regresión Lineal Simple. Cuando depende de más de una variable, es llamada Regresión Lineal Múltiple. En cambio, si $f(x)$ no es una función lineal, se dice que la regresión es no lineal.

2.3. Modelos lineales generalizados

Los MLG son una generalización de los modelos de regresión. Permiten relacionar la variable de respuesta Y que puede no seguir una distribución Normal con los predictores lineales $X's$ por medio de una función de enlace.

2.3.1. Componentes del modelo

La variable de respuesta Y en un MLG es explicada por una fuente de variabilidad de tipo aleatoria y otra de tipo determinista, relacionadas a través de una función de enlace.

1. *Componente aleatoria*: Una función de densidad o cuantía f perteneciente a la familia exponencial.

$$f(y) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)}\right\} c(y, \phi) \tag{2.5}$$

   Donde θ es el parámetro canónico o parámetro de la familia, que depende de los regresores a través de la función link, la que linealiza la relación entre la variable de respuesta Y y las variables explicativas $X's$, y ϕ es un parámetro de dispersión por lo general conocido. Las funciones $a(.)$, $b(.)$ y $c(.)$ son conocidas.

El hecho de que $f(y)$ pertenezca a la *familia exponencial* le otorga ciertas propiedades muy convenientes, como que se pueda encontrar fácilmente un estimador suficiente (que contiene toda la información de la muestra) e insesgado ($E(\hat{\theta}) = \theta$) para el parámetro poblacional (23). Además, la primer y segunda derivada con respecto a θ representan la media y la varianza de la variable respectivamente.

2. *Componente sistemático - Predictor lineal*: Especifica las variables explicativas y las relaciona al modelo por $X\beta$. β es el vector de parámetros de regresión y es estimado por Máxima Verosimilitud

3. *Función link*: Relaciona la distribución de la variable dependiente con las variables explicativas.

$$g(Y) = X\beta + \epsilon \tag{2.6}$$

De este modo, la función de enlace relaciona el Componente Aleatorio con el Componente Sistemático del modelo.

En ocasiones las observaciones de Y indican éxito o fracaso (*binaria*), o categorías, y se lleva a cabo una regresión Logística para el estudio de la población. En otros casos, como el del presente trabajo, cada observación es un conteo, por lo que se puede asignar a Y una distribución Poisson, Binomial Negativa, Poisson Inversa Gaussiana, Binomial Negativa-ρ.

2.3.2. Estimación de los parámetros

La estimación de los parámetros en el MLG se realiza por Máxima Verosimilitud.

$$L(\theta/y) = \prod f(y_i) \tag{2.7}$$

Maximizar el logaritmo de esta función es equivalente a maximizar dicha función.

$$log(L(\theta/y)) = \sum log(f(y_i)) = \sum \left(\frac{y_i\theta - b(\theta)}{a(\phi)} + log(c(y_i, \phi)) \right) \tag{2.8}$$

El estimador máximo verosímil de θ anula la derivada de la función anterior.

$$\frac{\delta logL}{\delta\theta} = \sum \frac{y_i - b'(\theta)}{a(\phi)} \tag{2.9}$$

Como estas ecuaciones de estimación no se pueden resolver directamente, su solución se aproxima por métodos iterativos, como el algoritmo de Newton-Raphson (N-R).

Algoritmo de Newton-Raphson. (26)

Dado un parámetro inicial estimado $\widehat{\theta}^0$ que puede ser estimado por el Método de los Momentos, podemos obtener una aproximación de L alrededor de $\widehat{\theta}^0$

$$L^*(\theta) = L(\widehat{\theta}^0) + L'(\widehat{\theta}^0)(\theta - \widehat{\theta}^0) + \tfrac{1}{2}(\theta - \widehat{\theta}^0)^2 H_n(\widehat{\theta}^0) \approx L(\theta)$$

Entonces podemos maximizar L^* alrededor de θ produciendo un nuevo valor del parámetro que llamaremos $\widehat{\theta}^1$. La condición para resolver este problema es:

$$L'(\widehat{\theta}^0) + H_n(\widehat{\theta}^0)(\widehat{\theta}^1 - \widehat{\theta}^0) = 0$$

que es lo mismo que:

$$\widehat{\theta}^1 = \widehat{\theta}^0 - [H_n(\widehat{\theta}^0)]^{-1} L'(\widehat{\theta}^0)$$

La regla general de iteración de N-R es:

$$\widehat{\theta}^{t+1} = \widehat{\theta}^t - [H_n(\widehat{\theta}^t)]^{-1} L'(\widehat{\theta}^t)$$

donde $H_n(.)$ es la matriz Hessiana.

El procedimiento iterativo termina cuando se satisface un criterio de convergencia predefinido que puede ser: el cambio en $\widehat{\theta}^{t+1} - \widehat{\theta}^{t}$, o el valor de $L'(\widehat{\theta}^{t})$. La convergencia ocurre cuando alguno de esos valores es cercano a cero.

2.4. Modelos de regresión para datos de conteo

Los datos de conteo son observaciones de valores enteros no negativos que comienzan en cero. Una variable de conteo es una lista específica de datos de conteo que toma valores no negativos y donde cada valor es independiente a otro. Una variable de conteo es aquella que determina el número de eventos que ocurren en un determinado espacio o tiempo. En este caso el modelo de regresión relaciona la variable de conteo a explicar Y, con una o más variables predictoras X que pueden ser categóricas o cuantitativas. La variable de respuesta Y no tiene límite superior y toma el valor cero en muchos casos.

El objetivo principal al modelar datos de conteo es explicar el número de ocurrencias de un evento en un momento o espacio determinado. Así, la variable a explicar Y toma valores enteros no negativos. Se estiman los parámetros de una distribución de probabilidad que se considera apropiada para representar los datos a modelar. Las distribuciones más utilizadas para representar datos de este tipo son: Poisson, Binomial Negativa (BN), Poisson Inversa Gaussiana (PIG) y Binomial Negativa-ρ (BN-ρ).

La distribución Poisson tiene un solo parámetro, μ, que es su media y varianza. La condición que deben cumplir los datos es que la media y la varianza deben ser iguales, por lo que a medida que aumenta el valor esperado de la variable Y, mayor variabilidad. Cuando se cumple la igualdad de media y varianza en la variable a explicar, se conoce como criterio de equidispersión, y no suele cumplirse al trabajar con datos reales. El método más usado para hacer frente a la sobredispersión de

Poisson es modelar los datos usando una Binomial Negativa. La distribución Binomial Negativa tiene un parámetro adicional llamado parámetro de dispersión; es una medida de ajuste para acomodar el exceso de variabilidad en los datos. Esta distribución permite mayor flexibilidad al modelar datos sobredispersos. Cuando los datos presentan gran concentración en los primeros valores del recorrido, es útil usar la distribución Poisson Inversa Gaussiana, que es una mezcla de una variable aleatoria Poisson donde su parámetro se distribuye de acuerdo a una distribución Inversa Gaussiana. El modelo BN-ρ, es un modelo de conteo de tres parámetros, donde ρ es el exponente del segundo término de la varianza, lo que da una mayor flexibilidad en la misma.

- Estructura del modelo:

  Dado que la función link es el logaritmo, el modelo tiene la forma

  $$\widehat{y} = \exp\{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k\} \Rightarrow log(\widehat{y}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

  La función logaritmo garantiza que los valores predichos serán siempre positivos.

- Tipos de Modelos de Conteo:

  Al elegir el modelo más apropiado para los datos, se está seleccionando una distribución de probabilidad o mezcla de distribuciones que mejor describen los datos de la población de los cuales se extrajo la muestra a ser modelada. Los datos no siempre se asocian a una distribución Poisson o Binomial Negativa. Puede suceder que no asuman valores cero o que tengan gran cantidad de ceros, por lo que es necesario un ajuste a la función de probabilidad. Con este propósito se usan Modelos Cero Truncados (MCT), Hurdle (MH) y Cero Inflados (MCI). Pertenecen a los *Modelos en dos Partes* ya que presentan un componente Logit o Probit para determinar los conteos cero frente a los conteos positivos, y un modelo Poisson, PIG o Binomial Negativo para los

conteos positivos.

- Estimación de los parámetros del modelo:

  Se estiman por Mínimos Cuadrados Generalizados o Máxima Verosimilitud Iterativos (Algoritmo de Newton-Raphson (26))

2.4.1. Regresión Poisson

El modelo regresión Poisson permite relacionar la variable aleatoria Y con distribución Poisson con variables explicativas X por medio de la función de enlace logaritmo.

Entonces el modelo de regresión Poisson es de la forma:

$$log(Y) = X\beta + \epsilon$$

La variable aleatoria Y con distribución Poisson es una variable discreta y es la más simple usada para modelar datos de conteo. Es unimodal y se destaca por la propiedad de igualdad de media y varianza, lo que lleva a que cuando el valor de los conteos aumenta en media, también aumenta en variabilidad.

Se aplica a los casos en que se busca modelar un número de ocurrencias de un evento o fenómeno de cierto tipo que se producen en un intervalo de tiempo o espacio de observación (7).

Su función de cuantía está dada por:

$$f(y;\lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^y}{y!} & y = 0, 1, 2, ... \\ 0 & \text{otro caso} \end{cases} \tag{2.10}$$

Donde y es el número de ocurrencias del evento o fenómeno y λ el número medio

de veces que se espera que ocurra el mismo en el intervalo de tiempo o el espacio de observación y coincide con la varianza.

Supuestos de la distribución Poisson:

1. Distribución discreta con un solo parámetro: $\lambda = media$
2. Y toma valores enteros no negativos
3. Las observaciones son independientes entre sí
4. La Media y la Varianza son iguales. A mayor media, más variabilidad

En la figura 2.1 se muestra cómo varía la distribución de Poisson a través de su función de cuantía según el valor que toma λ.

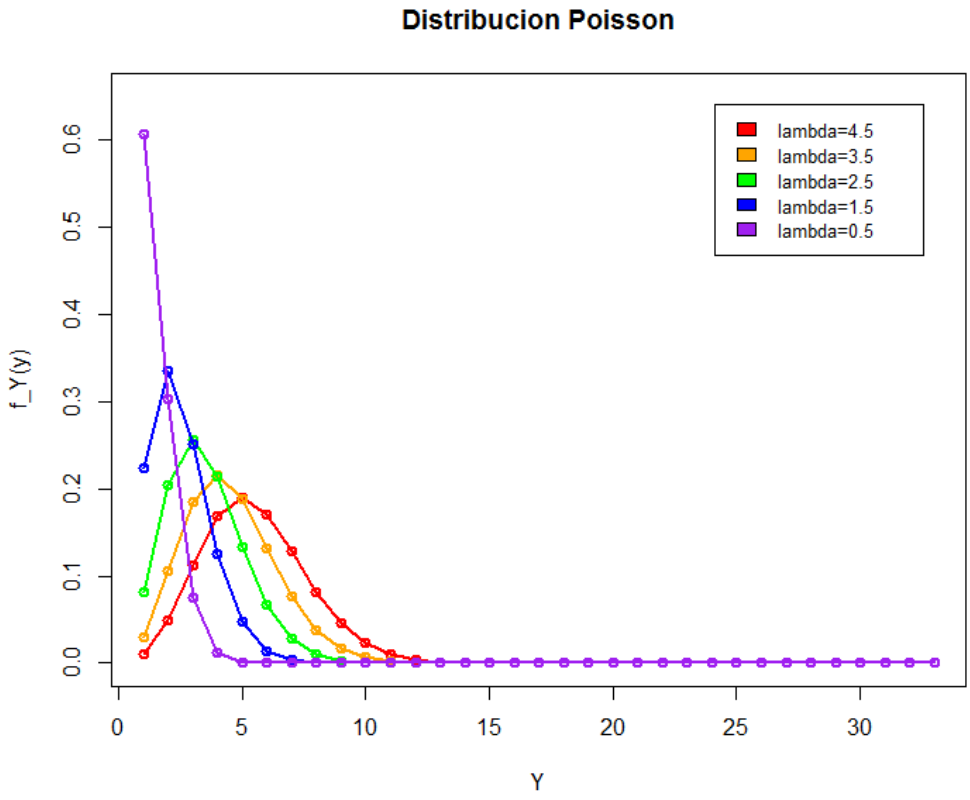

Figura 2.1: *Distribución Poisson según* λ

Esta v.a. pertenece a la familia exponencial: Su cuantía $f(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$ se puede

18

expresar:

$$f(y;\lambda) = \exp\{-\lambda + ylog\lambda\}\,\frac{1}{y!} \tag{2.11}$$

Donde:

$$\theta = log\lambda$$

$$b(\theta) = e^{-\theta}$$

$$c(y,\phi) = \tfrac{1}{y!}$$

$$\phi = 1$$

La esperanza y la varianza están dadas por:

$$E(y_i) = \mu_i = b'(\theta)$$

$$V(y_i) = \phi b"(\theta)$$

La función de enlace en este caso es $g(y) = log(y)$. $E(y) = V(y) = \lambda$ y el parámetro de dispersión $\phi = 1$.

La interpretación de los coeficientes estimados β_j, dada la forma del modelo $log(\hat{y}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$, debe verse como el cambio en el logaritmo del valor esperado de la variable de respuesta y cuando cambia en una unidad la variable predictora x_j manteniendo las otras constantes.

La propiedad de equidispersión muy frecuentemente no se cumple al trabajar con datos reales; la hipótesis de igualdad de media y varianza no suele ser cierta, dado que la varianza observada por lo general es mayor que la media observada (sobredispersión). Una forma de hacer frente a este problema es dejar el parámetro de dispersión ϕ libre. Este es el caso de la distribución Quasi-Poisson, que usa la función de me-

dia y varianza de la Poisson pero deja irrestricto el parámetro de dispersión. Otro problema que suele surgir al modelar datos reales es el exceso de ceros, para lo que se presentan distintas soluciones más adelante.

Sobredispersión en modelos de regresión Poisson

Cuando se trabaja con una base de datos de conteo puede ocurrir que el modelo Poisson puede parecer sobredisperso y en realidad no lo es, o puede efectivamente presentar sobredispersión.

En el primer caso, una simple corrección al modelo puede hacer desaparecer la variabilidad no deseada. Si luego de los ajustes el problema de sobredispersión no desaparece, se presenta el segundo caso, donde se deben buscar modelos alternativos, que puede ser el modelado a partir de la distribución Binomial Negativa (no tiene restricción de igualdad en media y varianza), o el uso de modelos más complejos como son los modelos compuestos y modelos en 2 partes.

Si existe evidencia suficiente para probar que los datos no siguen una distribución Poisson, entonces será necesario emplear un modelo de conteo alternativo que se ajuste al tipo de supuesto violado en la distribución de los datos; por ejemplo:

- MCT Poisson para el caso en que los datos no admiten el conteo cero
- MCI Poisson si hay más valores ceros de los esperados para una distribución de Poisson para una media dada o los conteos cero provienen de una fuente diferente que los conteos mayores que cero. Los conteos cero se admiten en ambos componentes del modelo.
- MH si hay más valores cero o menos valores cero basados en la distribución Poisson para una media dada o los conteos cero provienen de una fuente diferente que los conteos mayores que cero. En este modelo, los conteos cero se admiten sólo en el componente binario, mientras que el componente truncado

en cero no presenta ese valor en el recorrido.

- versiones con la Binomial Negativa de los puntos anteriores.

2.4.2. Regresión Binomial Negativa (6)

EL modelo regresión Binomial Negativa permite relacionar la variable aleatoria Y con distribución Binomial Negativa con variables explicativas X por medio de la función de enlace logaritmo.

Entonces el modelo de regresión Binomial Negativa tiene la forma:

$$log(Y) = X\beta + \epsilon$$

La variable aletoria Y Binomial Negativa es una variable discreta perteneciente a la familia exponencial que cuenta el número de fracasos antes del r-ésimo éxito en $Y+r$ experimentos independientes Bernoulli, siendo la probabilidad de éxito en cada prueba p. Un experimento de Bernoulli es tal que sólo admite dos posibles resultados: éxito o fracaso. La cantidad de pruebas es indefinida y sólo concluirá cuando se obtengan r resultados favorables.

Su función de cuantía está dada por:

$$P(Y = y; p, r) = \begin{cases} \binom{y+r-1}{y}(p)^r(1-p)^y & y = 0, 1, 2, ... \\ 0 & \text{otro caso} \end{cases} \tag{2.12}$$

Con $E(Y) = \frac{r(1-p)}{p}$ y $V(Y) = r\frac{(1-p)}{p^2}$

Supuestos del modelo Binomial Negativo:

1. La variable de respuesta es discreta y toma valores enteros no negativos.

2. A medida que la media μ aumenta, la probabilidad de un conteo=0 decrece.

3. El valor 0 se encuentra en el recorrido de Y.

4. La $V(Y)$ es mayor que la $E(Y)$.

La distribución Binomial Negativa se puede ver como una mezcla de distribuciones Poisson-Gamma: es una v.a. con distribución Poisson en la cual su parámetro es una variable aleatoria que se distribuye Gamma.

Sea la función de cuantía de la distribución Poisson

$$P(Y = y/\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}\mathbb{I}_{\{y=0,1,\ldots,n\}} \tag{2.13}$$

y la función de densidad de la v.a. Gamma

$$g(\lambda) = \tfrac{\alpha^\beta}{\Gamma(\beta)}\lambda^{\beta-1}e^{-\alpha\lambda} \quad \lambda \geq 0, \alpha > 0, \beta > 0 \tag{2.14}$$

$$\Rightarrow P(Y = y, \Lambda = \lambda) = P(Y/\lambda)g(\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}\frac{\alpha^\beta}{\Gamma(\beta)}\lambda^{\beta-1}e^{-\alpha\lambda} \tag{2.15}$$

$$\Rightarrow P(Y = y) = \int P(Y;\lambda)g(\lambda)d\lambda = \binom{y+\beta-1}{y}\left(\frac{\alpha}{\alpha+1}\right)^\beta\left(\frac{1}{\alpha+1}\right)^y \tag{2.16}$$

para valores de $y \geq 0$

Donde $\alpha = \frac{p}{1-p}$ y $\beta = r$, por lo que $E(Y) = \frac{\beta}{\alpha}$ y $V(Y) = \beta\frac{(\alpha+1)}{\alpha^2}$

Al ser $V(Y) > E(Y)$ esta distribución permite, a diferencia de la distribución Pois-

son, modelar datos sobredispersos, y suele ser la primera alternativa para hacer frente a la restricción de igualdad de media y varianza de dicha distribución.

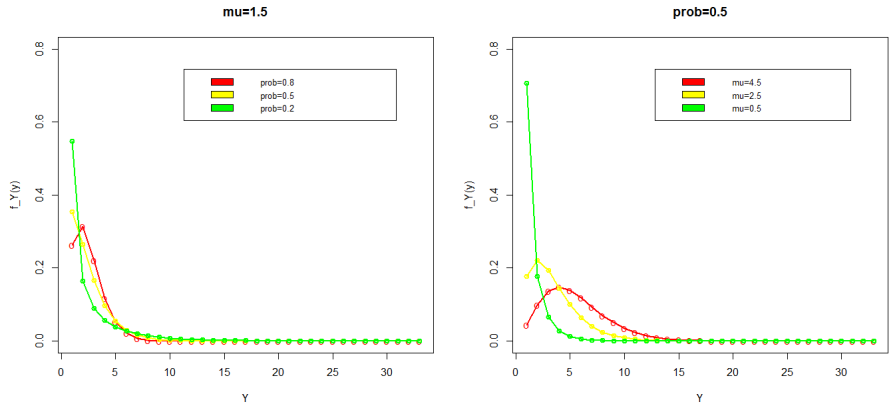

Figura 2.2: *Distribución Binomial Negativa con un parámetro fijo*

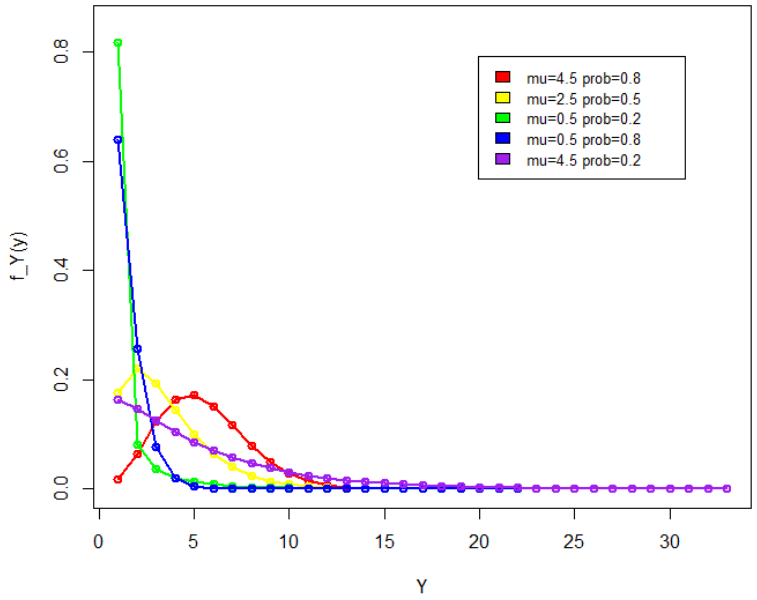

Figura 2.3: *Distribución Binomial Negativa según distintos parámetros*

En la figura 2.2 se muestra la forma que toma la cuantía de la distribución Binomial Negativa por un lado variando el parámetro p y dejando μ fijo (izquierda) y por el otro variando el parámetro μ y dejando p fijo (derecha). En la figura 2.3 se muestra la forma que toma la distribución Binomial Negativa también a través de su cuantía

variando ambos parámetros.

Parametrizaciones de la Varianza

Se puede distinguir hasta 13 formas de modelos Binomial Negativo, debido a las formas en que la varianza varía. Aquí se hará referencia a la forma lineal (BN1), la que puede ser vista como un modelo Quasi Poisson con $\phi = (1 + \alpha)$, la forma cuadrática y más tradicional (BN2) y la BN-ρ, donde la varianza varía en cada observación. La principal diferencia entre estos modelos radica en el valor que toma el exponente en la función de varianza, lo que se muestra en la tabla 2.1.

Tabla 2.1: *Media y Varianza de los distintos tipos de distribución Binomial Negativa*

Modelo	**Media**	**Varianza**
BN1	μ	$\mu(1+\alpha) = \mu + \alpha$
BN2	μ	$\mu(1+\alpha\mu) = \mu + \alpha\mu^2$
BN-ρ	μ	$\mu(1+\alpha\mu^{\rho-1}) = \mu + \alpha\mu^{\rho}$

BINOMIAL NEGATIVA-ρ: Este modelo tiene un parámetro adicional en el segundo término en la varianza.

$$V(Y) = \mu + \alpha\mu^{\rho}$$

Siendo $\alpha = \frac{1}{r}$

La estimación de ρ nos indica si es conveniente modelar los datos a partir de una BN1 o con BN2, y la elección de la distribución se realiza mediante el Test de Razón de Verosimilitud.

2.4.3. Otras formas de tratar la sobredispersión: Regresión Poisson Inversa Gaussiana (PIG)

Al igual que la distribución Binomial Negativa, la distribución Poisson Inversa Gaussiana es una mezcla de distribuciones. La variable aleatoria Y sigue una función de cuantía Poisson (2.10) donde su parámetro λ es también una variable aleatoria con distribución IG.

Debido a la flexibilidad de esta distribución, la distribución Poisson Inversa Gaussiana tiene la capacidad de modelar datos de conteo con alta sobredispersión.

Es una alternativa al modelo Binomial Negativo cuando se trata de ajustar datos sobredispersos. Además, es útil para modelar datos que tienen gran concentración en los primeros valores, y esto es una ventaja sobre la BN.

2.5. Exceso de ceros en datos de conteo

Las distribuciones que fueron tratadas anteriormente asumen que pueden existir datos iguales a cero. Algunas variables de conteo que describen datos reales muestran un porcentaje de ceros muy alto. Esa cantidad de ceros no es compatible con las distribuciones Poisson o BN. La gran diferencia entre el número esperado y el número observado de ceros es un problema en el análisis: puede ser causa de sobredispersión y la estimación de los coeficientes puede no ser fiable. Subestima la varianza con intervalos de confianza más chicos de lo que corresponde, obteniendo como consecuencia variables significativas que no lo son. Además, la precisión en las inferencias se verán altamente afectadas. Para corregir este problema se debe hacer un ajuste a la función o usar otro modelo diferente.

2.5.1. Modelos truncados en cero

Los modelos truncados implican que en algún punto del recorrido de la variable, un determinado valor está totalmente ausente.

Si el valor que no se observa es el cero entonces se dice que es un modelo "Truncado en Cero". Este tipo de modelos no admite conteos ceros, por lo que la distribución no debe tener este valor en su recorrido para poder modelar los datos adecuadamente. Es necesario modificar la función para que la suma de las probabilidades de los valores sea 1.

Las distribuciones presentadas anteriormente pueden ser modificadas para llegar a sus versiones truncadas.

- Poisson Cero Truncado: Como en una distribución Poisson la $P(Y = 0) = e^{-\lambda}$ y para valores de la media cada vez más grandes la $P(Y = 0)$ es cada vez más chica (dado que a mayores valores de λ, $e^{-\lambda}$ es cada vez más chico), usar este modelo no es necesario si la media es alta (por ejemplo mayor que 5) En este caso la distribución observada es de la forma

$$f(y/x, y > 0) = \frac{f(y, y > 0/x)}{f(y > 0/x)} = \frac{exp(-\lambda)\lambda^y}{y!(1 - exp(-\lambda))} \tag{2.17}$$

  Para $y > 0$, donde y son los valores observados (en este caso mayores que 0) y x son las variables explicativas.

- Binomial Negativa Cero Truncado: La lógica es la misma que para la distribución Poisson, se trunca la distribución en $y = 0$ y la distribución observada es:

$$f(y/x, y > 0) = \frac{\binom{y + \beta - 1}{y} \left(\frac{\alpha}{\alpha+1}\right)^{\beta} \left(\frac{1}{\alpha+1}\right)^{y}}{1 - \left(\frac{\alpha}{\alpha-1}\right)^{\beta}} \tag{2.18}$$

- Poisson Inversa Gaussiana Cero Truncado: Como se mencionó, el modelo PIG permite trabajar con datos sobredispersos. Es ideal para trabajar con datos asimétricos y que no permiten conteo cero, aunque no ajusta tan bien como la Binomial Negativa.

- Binomial Negativa-ρ Cero Truncado: Es el mejor modelo ajustado si hay heterogeneidad en la dispersión.

2.5.2. Modelos de regresión Hurdle

El Modelo de Regresión Hurdle, también conocido como "Modelo con Obstáculo", es un modelo de dos componentes o "modelo en dos partes"que combina:

1. Un proceso binario para los valores que están por encima o por debajo del valor de selección, modelado por medio de un proceso logit, para describir la probabilidad de que se cruce el "obstáculo". Dicho proceso modela datos que toman dos valores: éxito o fracaso. Este componente del modelo sólo genera conteos cero.

   Sea y_i la observación i

   $y_i \sim Ber(p_i)$ siendo $p_i = E(y_i/x_i)$ la probabilidad de éxito.

   El modelo logístico es (16)

$$E(Y/X) = \pi_i = \frac{e^{X\beta}}{1 + e^{X\beta}} \tag{2.19}$$

$$\pi_i = \frac{1}{1 + e^{-X\beta}} \tag{2.20}$$

   Haciendo cálculos se llega a

$$\frac{\pi_i}{1-\pi_i} = \frac{1+e^{X\beta}}{1+e^{-X\beta}} = e^{X\beta} \tag{2.21}$$

Aplicando logaritmo en ambos lados de la ecuación se obtiene:

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = X\beta \tag{2.22}$$

lo que se conoce como transformación *logit* de π_i.

$\frac{\pi_i}{1-\pi_i}$ es conocido como *odds*, que es una razón de probabilidades: es el cociente entre la probabilidad de que ocurra el evento y que no ocurra el evento, por lo que cuanto más alto el *odds*, más alta será la probabilidad de que el suceso ocurra.

2. Un proceso que genera sólo los conteos mayores que cero mediante un modelo Cero Truncado. Este componente se puede modelar mediante un modelo Poisson, Binomial Negativo o PIG.

El modelo Hurdle tiene la forma

$$f_{hurdle}(y; x, z, \beta, \gamma) = \begin{cases} f_{cero}(0; z, \gamma) & \text{si } y = 0 \\ (1 - f_{cero}(0; z, \gamma)) f_{cont}(y; x, \beta)/(1 - f_{cont}(0; x, \beta)) & \text{si } y > 0 \end{cases} \tag{2.23}$$

En este modelo, se considera que los datos son generados de tal forma que un proceso genera conteos positivos luego de cruzar un obstáculo. Hasta que dicha barrera es cruzada, el proceso genera conteos cero. El vector de parámetros β y γ del modelo se estiman por máxima verosimilitud y pueden ser maximizados por separado.

2.5.3. Modelos de regresión Cero Inflado

El modelo de regresión cero inflado fue propuesto por Lambert (10) con el fin de, como en los modelos Hurdle, tratar el problema en los datos que muestran una

cantidad de ceros mucho más alta que la que es compatible con las distribuciones usualmente utilizadas, es decir, los datos a ser estudiados presentan más ceros que los esperados. Al igual que los modelos de regresión Hurdle, este modelo combina una variable binaria con un modelo de conteo Poisson, PIG o Binomial Negativo. Es un modelo mixto de dos componentes que da mayor peso a la probabilidad de que la variable sea igual a cero, por lo que la función de probabilidad para un modelo de regresión Cero Inflado es una mezcla de una función de masa concentrada en cero y un modelo perteneciente a la familia exponencial. A diferencia de los Modelos Hurdle, el primer componente genera sólo conteos cero, pero el segundo genera el rango completo de conteos, incluyendo los ceros.

El MCI tiene la forma:

$$f_{ceroinf}(y; x, z, \beta, \gamma) = \begin{cases} f_{cero}(0; z, \gamma) + (1 - f_{cero}(0; z, \gamma)) f_{cont}(0; x, \beta) & \text{si } y = 0 \\ (1 - f_{cero}(0; z, \gamma)) f_{cont}(y; x, \beta) & \text{si } y > 0 \end{cases} \tag{2.24}$$

De esta forma, se generan dos modelos y luego se combinan.

Un signo positivo en la estimación del coeficiente del componente binario indica que, si la variable toma el valor de referencia, la probabilidad de un conteo mayor que cero aumenta. En cambio para el componente de conteos, la interpretación de los parámetros debe hacerse de igual forma que en modelos Poisson y Binomial Negativo.

2.6. Evaluación del ajuste

Una vez elegido el modelo de regresión es necesario evaluar si el mismo tiene un buen ajuste y si es el indicado para los mismos. Eso implica analizar errores y realizar

tests para corroborar la bondad del ajuste y la elección del modelo. Además, citando a Joseph M. Hilbe (6) en su libro *Negative Binomial Regression*, "Un modelo sólo es tan bueno como los resultados de sus ajustes estadísticos".

2.6.1. Análisis de los residuos (2)

Los residuos son definidos como $r_i = y_i - \widehat{y_i}$, $i = 1, ..., n$.

Esta medida parte de la diferencia entre el valor ajustado y el valor observado de la variable dependiente.

2.6.2. Test pseudo R^2

El estadístico R^2 es una herramienta para analizar los modelos de regresión ordinarios, y es conocido como Coeficiente de Determinación, indicando el mismo el porcentaje de variación en los datos que es explicado por el modelo. El estadístico toma valores de 0 a 1, siendo 1 el mejor ajuste del modelo. Este estadístico no es apropiado para evaluar modelos de regresión no lineales, como es el caso de los MLG. En este caso, el estadístico usado es Pseudo R^2, que también varía entre 0 y 1 y es definido como:

$$R_P^2 = 1 - L_F/L_I \tag{2.25}$$

donde L_F es la log-verosimilitud del modelo ajustado con las variables explicativas y L_I es la log-verosimilitud del modelo sólo con la intercepción y sin variables explicativas.

Al comparar modelos, los modelos con valores de R_P^2 más bajos, indican un ajuste más "pobre", ya que tienen una menor verosimilitud, la cual lleva a un menor R_P^2.

2.6.3. Test de bondad de ajuste del desvío

El Desvío es expresado como

$$D = 2\sum_{i=1}^{n}\{L(y_i; y_i) - L(\beta_i; y_i)\} \tag{2.26}$$

donde $L(y_i; y_i)$ es la log-verosimilitud del modelo saturado, donde cada valor de μ es reemplazado por el valor de cada y_i dado, y $L(\mu_i; y_i)$ es la log-verosimilitud del modelo a ser estimado. Es un test cuyo estadístico tiene distribución Chi2, donde los grados de libertad es el número de predictores del modelo incluída la intercepción. Si el $p - valor$ resultante del valor del $Chi2$ es menor que el nivel de significación, entonces se rechaza la hipótesis nula.

$H_0) D = 0$

$H_1) D > 0$

2.6.4. Test razón de verosimilitud.

Este test compara modelos con algunos predictores contra el mismo modelo con más predictores. Evalúa si las variables explicativas deben mantenerse en el modelo, es decir, si tienen información para explicar el comportamiento de la variable y.

$$LR = -2(L_R - L_F) \tag{2.27}$$

Donde L_R es la verosimilitud del modelo reducido y L_F la del modelo más completo.

2.6.5. Criterios de selección del modelo

Los tests de criterios de selección del modelo son tests comparativos, siendo los que presentan valores menores los que indican un mejor ajuste. Los principales tests de Criterio de la Información son Akaike Information Criterion (AIC) y Bayesian Information Criterion (BIC). Estos criterios consisten en una serie de parametrizaciones alternativas, cada una de las cuales tiene como objetivo determinar un método para evaluar mejor el ajuste del modelo.

- Criterio de Información de Akaike (AIC): el estadístico AIC tiene la forma:

$$AIC = \frac{-2(L-k)}{n} \tag{2.28}$$

  donde L representa la verosimilitud del modelo, k el número de predictores y n el número de observaciones.

  $2k$ penaliza la cantidad de predictores, dado que al aumentar la cantidad de los mismos el modelo es más verosímil entonces $-2L$ se vuelve más chico. Por el principio de parsimonia, en igualdad de condiciones, el modelo más sencillo, suele ser el mejor.

- Criterio de Información Bayesiana (BIC): el estadístico BIC tiene la forma:

$$BIC = -2L + klog(n) \tag{2.29}$$

  donde L representa la verosimilitud del modelo, k el número de predictores y n el número de observaciones.

  Este estadístico da un mayor peso al término de ajuste $klog(n)$ que el AIC.

Capítulo 3

Datos de la aplicación

En el año 1935 H. Klein, C. E. Palmer, y J. W. Knutson desarrollaron el llamado índice CPO (9) (por la primer letra de las palabras Cariado-Perdido-Obturado), con el fin de estudiar el estado dental de niños de algunas ciudades norteamericanas. El mismo se ha convertido en el índice fundamental para los estudios odontológicos al momento de cuantificar la existencia de caries dental, ya que tiene en cuenta la existencia de caries tanto presente como pasada. Así el índice considera:

C: Cariado - Enfermedad presente: la lesión está activa

P: Perdido - Enfermedad pasada: la pieza fue perdida por caries dental

O: Obturado - Enfermedad curada: la pieza recibió tratamiento.

Según la Revista de Salud "Índices Epidemiológicos Para Medir La Caries Dental" (M. Fernández PratTs) el Índice CPO es un "Índice fundamental de los estudios odontológicos que se realizan para cuantificar la prevalencia de la caries dental. Señala la experiencia de caries tanto presente como pasada, pues toma en cuenta los dientes con lesiones de caries y con tratamientos previamente realizados. Se obtiene de la sumatoria de los dientes permanentes cariados, perdidos y obturados, incluidas las extracciones indicadas, entre el total de individuos examinados."

En Uruguay existe el llamado Programa Nacional de Salud Bucal que propone "Contribuir al logro del más alto grado posible de salud bucal de la población uruguaya, impulsando, promoviendo y articulando las adecuadas acciones promocionales, preventivas y asistenciales integradas en un Sistema de Salud y que correspondan a las necesidades de cada individuo" (1) y que forma parte del Sistema Nacional Integrado de Salud ($SNIS$), donde los grupos prioritarios son las mujeres embarazadas, los niños y los adolescentes, no existiendo cobertura para la población adulta. Al no contar con información de dicha población se realiza un relevamiento epidemiológico durante los años 2010-2011. Es el primer estudio de este tipo realizado en el país, basado en la metodología propuesta por la OMS que fue llevado a cabo por la Facultad de Odontología de la Universidad de la República (UDELAR), auspiciado por el Ministerio de Salud Pública (MSP).

Para el relevamiento de datos se consideraron 2 características principales de Uruguay: la concentración de la población en Montevideo y la distribución de las rutas nacionales en abanico desde Montevideo hacia el interior del país. Es un estudio realizado a la población joven y adulta urbana en sus domicilios, cuyo muestreo fue en 2 fases, en la primer fase de el conjunto de personas de los tres tramos etarios pertenecientes a localidades de 20.0000 o más habitantes de la Encuesta Contínua de Hogares (ECH) y en la segunda fase se seleccionan personas de la primera fase hasta llegar al tamaño de muestreo previamente calculado (tabla 3.1). El diseño muestral no fue tomado en cuenta en este trabajo.

Se aplicó un cuestionario a una muestra representativa de la población joven y adulta de todo el país, sobre la condición socio-económica, hábitos y factores de riesgo, utilización de servicios de salud y autopercepción de salud. Debido a que es el primer estudio de este tipo a nivel nacional, los examinadores fueron mayormente docentes de la Facultad de Odontología.

La información recogida refleja los principales problemas de salud bucal y las nece-

sidades de tratamiento en los grupos de edades de 15 a 24, 35 a 44 y 65 a 74 años, por medio de un examen bucal a partir de los criterios de la OMS

Tabla 3.1: *Proporción de personas relevadas por Región según Tramo Etario*

Tramo Etario	Montevideo	Interior	Total
15-24	0,50	0,45	**0,47**
35-44	0,23	0,25	**0,24**
65-74	0,27	0,30	**0,29**

3.1. Variables relevadas

Se aplica un cuestionario a las personas de la muestra seleccionadas referente a datos personales y demográficos, características socioeconómicas, acceso a servicios de salud, hábitos de riesgo y enfermedades generales. Luego de la aplicación del cuestionario se realiza un examen bucal por parte del examinador, donde se observan: lesiones de mucosa y pérdida dentaria, condición periodontal y pérdida de inserción en los tramos etarios de 35-44 y 65-74 años, caries dental en corona y raíz (este último en los tramos etarios de 35-44 y 65-74 años) y maloclusiones (en el grupo de edad de 15-24 años).

En esta sección se realiza un primer análisis a los datos obtenidos a partir de la aplicación de dicho cuestionario.

3.1.1. Variables a explicar: CPO, C, P y O

Para recoger la información necesaria sobre las enfermedades bucales consideradas, el examinador realizó un examen bucal en el domicilio a cada individuo.

El Índice CPO se calculó de acuerdo a lo recomendado por la OMS de la siguiente

Tabla 3.2: *Cantidad de personas por variable según tramo de prevalencia*

Valor	Ccorona	Pcorona	Ocorona	CPOcorona
0-5	1381	860	1239	548
6-10	75	152	182	225
11-15	16	91	50	174
16-20	7	85	9	122
21-25	0	98	1	150
26-32	0	196	1	260
NA's	6	3	3	6
Total	1485	1485	1485	1485

forma: se considera C como lesión de caries y diente obturado y cariado, P es el diente perdido por caries para todas las edades y para las personas mayores de 35 se consideran los dientes perdidos por otra razón, y para el componente O se consideran los dientes obturados sanos. En el presente trabajo sólo se tuvo en cuenta el componente "Corona"de cada diente, esto significa evaluar la parte visible de la pieza a diferencia de la raíz, por lo que las variables a estudiar serán: "Ccorona", "Pcorona", "Ocorona" y "CPOcorona". Los valores que presentan esas variables fueron separados en tramos de a 5 y se presenta en la tabla 3.2

Se calculó el índice de Knutson para la proporción de individuos libres de caries, individualizándose la proporción de sujetos con índice CPO igual a cero. El índice de Knutson discrimina entre el porcentaje de personas que presentan caries y las que no(18).

Se recogió información sobre otras enfermedades como paradenciopatías, lesiones de mucosa y maloclusiones, pero no se tomarán en cuenta en el presente estudio.

3.1.2. Características demográficas y socioeconómicas utilizadas en este trabajo

Los datos personales permiten ubicar al individuo en tramo etario, sexo, región y si tienen estudios universitarios o no. Para la clasificación socioeconómica se utiliza el

Índice de Nivel Socio Económico (INSE) elaborado por los economistas Fernández y Perera en el año 2003: Indice de Niveles Socieoconómicos (INSE) (4) y que fue validado por la Facultad de Ciencias Sociales de la Universidad de la República a través del Departamento de Sociología, y actualizado por las economistas Llambí y Piñeyro en el año 2012. El mismo toma valores de 0 a 100.

El INSE en su versión reducida, que se utilizó para el cuestionario, considera 9 variables referidas a características de la vivienda, servicios y tenencia de bienes (servicio doméstico en el hogar, heladera con freezer, TV color, automóvil, tarjetas de crédito internacional, número de baños en la vivienda), características de los miembros del hogar (ocupación del jefe del hogar, nivel educativo) y características de los ingresos del hogar (número de preceptores de ingreso).

Los valores que toman las variables del tipo socioeconómico y demográfico se muestran en la tabla 3.3

Tabla 3.3: *Proporción de personas por tramo etario, sexo, región, estudio universitario e INSE*

Tramo Etario		**Sexo**		**Región**		**Estudio Universitario**		**INSE**	
15 a 24	0,48	F	0,57	Interior	0,62	Si	0,27	Mínimo	0
35 a 44	0,24	M	0,43	Montevideo	0,38	No	0,72	Media	36,42
65 a 74	0,28					NA's	0,01	Máximo	89

3.1.3. Factores de riesgo

En la tabla 3.4 se muestran los valores que toman los factores de riesgo tomados en cuenta en el presente estudio, es decir, el consumo de mate y de tabaco del individuo encuestado. Se incluyen en el cuestionario el consumo de alcohol y consumo de frutas y verduras pero no fueron tomadas en cuenta en el presente trabajo.

Tabla 3.4: *Proporción de personas en la muestra según consuma o no mate o tabaco*

Consume Mate	**Fuma**
Si 0,75	0,25
No 0,24	0,74
NA's 0,01	0,01

3.1.4. Atención a la salud

Con referencia al acceso del encuestado a los servicios de salud, se toma en cuenta si el individuo cuenta con Institución Médica Colectiva, lo que se refleja en la siguiente tabla.

Tabla 3.5: *Proporción de personas en la muestra según tenga o no institución medica colectiva*

Institución Médica Colectiva	
Si	0,56
No	0,44
NA's	≈ 0

Capítulo 4

Resultados

El análisis computacional de este trabajo se realizó mediante el software libre R (19). Las librerías usadas fueron *pscl* (8), *sandwich* (28), *lmtest* (30), *MASS* (25), *gamlss* (21), *vcd* (15), *VGAM* (27) y *rcompanion* (14).

La variable CPO fue analizada en primer lugar a través de cada uno de sus componentes, C, P y O. Luego se analiza la variable CPO propiamente dicha, seleccionando las posibles familias de distribuciones que se ajusten a las mismas para luego estimar un modelo de regresión que describa cada una de éstas.

4.1. Variable Ccorona (Caries de corona)

Esta variable representa el número de dientes cariados en su corona, o sea el número de dientes con enfermedad presente. Luego de un análisis descriptivo donde se evalúan distribuciones candidatas a la variable para encontrar las que mejor se adapten a los datos, se ajustan, con éstas, modelos de regresión.

Se trabaja con datos sin valores faltantes, por lo que para la variable Ccorona se tiene un total de 1466 individuos. Esta variable toma valores de 0 a 18, su media es

1.45 y su varianza 6.39, o sea que su varianza es 4.41 veces su media. En la figura 4.1 se puede ver que los datos se encuentran concentrados en el valor cero decreciendo hacia el valor máximo de la variable.

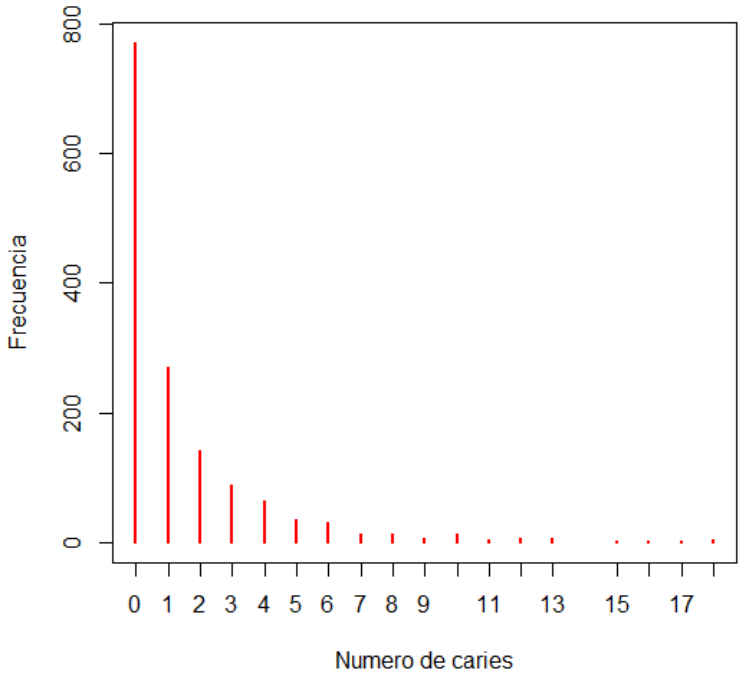

Figura 4.1: *Gráfico de Frecuencias de la variable Ccorona*

4.1.1. Distribución de Probabilidad para Ccorona

Para decidir cuales son las distribuciones que mejor se adaptan a la variable de interés se prueba el ajuste de las diferentes distribuciones que podrían adecuarse a los datos dadas sus características básicas.

Al tratarse de una variable cuantitativa discreta con recorrido no negativo, ya que es una variable de conteo, se intentarán ajustar las distribuciones Poisson y Binomial Negativa en primera instancia. Para una primera elección de candidata a distribución se compara gráficamente la distribución empírica con la distribución teórica estimada.

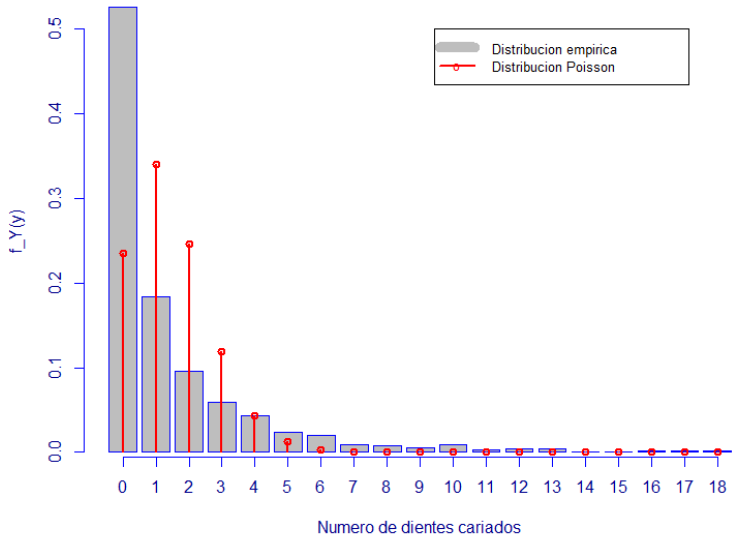

Figura 4.2: *Ajuste Poisson a la Variable Ccorona*

En la figura 4.2 se puede ver representada la función de distribución Poisson que mejor se ajusta a los datos a través de líneas rojas y la distribución empírica de Ccorona en barras azules. La distribución Poisson subestima la probabilidad del primer valor del recorrido de la variable así como sobreestima los 3 valores siguientes.

El número de conteos 0 estimado es considerablemente menor que el número real de conteo de ceros.

Por lo expuesto se prueba el ajuste de la distribución Binomial Negativa, que permite una varianza mayor a la media.

En la figura 4.3 se puede ver que esta distribución se ajusta mejor a la variable Ccorona, por lo que se podría preferir ésta a la distribución Poisson.

En este caso la estimación es muy similar a los valores reales para todos los valores que tomó la variable. Los parámetros de la distribución Binomial Negativa estimados según la notación de la ecuación (2.32) son $E(Ccorona) = \mu = 1{,}45$ y $r = 0{,}44$ que representa la cantidad de éxitos en $y + \beta$ experimentos, por lo que la varianza estimada es $V(Ccorona) = 6{,}24$ muy próxima a la varianza muestral.

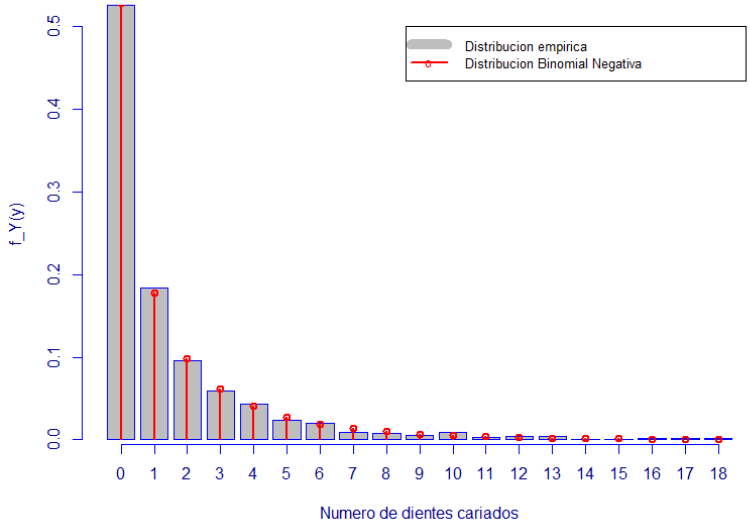

Figura 4.3: *Ajuste Binomial Negativa a la Variable Ccorona*

De las dos distribuciones de conteo analizadas la distribución Binomial Negativa es la que mejor se ajusta a los valores observados de la variable Ccorona, por lo que se estimarán modelos de regresión donde se asumirá que la variable Ccorona se distribuye Binomial Negativo.

4.1.2. Modelos de Regresión para Ccorona

Se quiere modelar la variable a explicar Ccorona con las variables explicativas región, tramo etario, sexo, estudio universitario, institución médica colectiva, consume mate, fuma, INSE.

A través de la figura 4.4 y de las tablas de relaciones bivariadas se muestra una visión general de la relación bivariada de cada una de las variables explicativas con la variable Ccorona viendo así las relaciones parciales.

Se puede notar que el número de personas que no tienen caries es mayor si tienen estudios universitarios que si no lo tienen, así como el 75 % de las personas que tienen estudios universitarios tienen 1 o menos caries, y el 75 % de las que no tienen estudio universitario tienen 2 o menos caries, esto es las personas que no tienen estudios universitarios tienen un número mayor de caries. Lo mismo sucede con las

personas según si tienen institución médica colectiva, el 75 % de las personas que tienen institución médica colectiva tienen 1 o menos caries, y el 75 % de las que no tienen institución médica colectiva tienen 3 o menos caries. El número de personas que no tienen caries es mayor en el grupo de las que no fuman que dentro de las que sí fuman, el 75 % de las personas que fuman tienen 3 o menos caries y el 75 % de las personas que no fuman tienen 1 o menos caries, o sea que las personas que fuman tienen más caries. Lo mismo sucede con la variable consume mate, el 75 % de las personas que consumen mate tienen 2 o menos caries y el 75 % de las personas que no consumen mate tienen 1 o menos caries. Para valores de INSE mayores a 20, a medida que aumenta el valor de esta variable, disminuye la cantidad de coronas con caries.

La cantidad de personas con caries o sin éstas no parece diferenciarse según si es hombre o mujer. Así como tampoco en los tramos de edad de 15 a 24 y de 35 a 44, aunque si pertenecen al tramo de 65 a 74 se puede ver una mayor cantidad de individuos sin caries y con hasta una caries lo cual es el 75 % de los individuos del grupo. La cantidad de personas sin caries es mayor dentro del grupo perteneciente a Montevideo que dentro del grupo del interior del país, aunque el 75 % de los individuos, tanto de Montevideo como del interior, tienen menos de 2 caries.

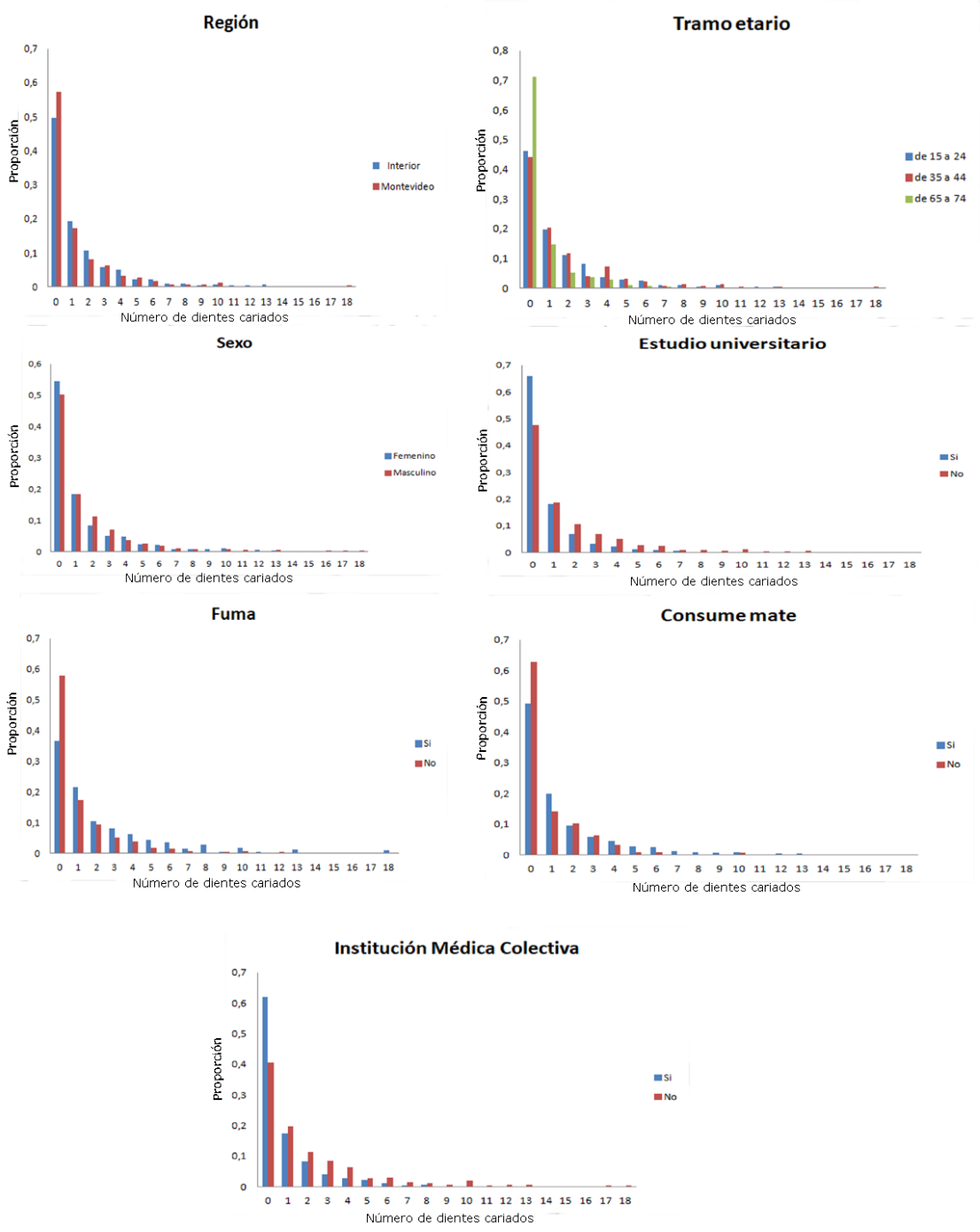
Región
Proporción
Número de dientes cariados
Interior
Montevideo
Tramo etario
de 15 a 24
de 35 a 44
de 65 a 74
Sexo
Femenino
Masculino
Estudio universitario
Si
No
Fuma
Consume mate
Institución Médica Colectiva

Figura 4.4

44

Tabla 4.1: *Medidas de resumen de Ccorona según Región*

Región	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Interior	0	0	1	1,57	2	18
Montevideo	0	0	0	1,24	2	18

Tabla 4.2: *Medidas de resumen de Ccorona según Tramo Etario*

Tramo Etario	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
15-24	0	0	1	1,68	2	18
35-44	0	0	1	1,89	2	18
65-74	0	0	0	0,65	1	11

Tabla 4.3: *Medidas de resumen de Ccorona según Sexo*

Sexo	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Femenino	0	0	0	1,36	2	18
Masculino	0	0	0	1,56	2	18

Tabla 4.4: *Medidas de resumen de Ccorona según Estudio Universitario*

Estudio Universitario	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Si	0	0	0	0,73	1	10
No	0	0	1	1,72	2	18

Tabla 4.5: *Medidas de resumen de Ccorona según Institución Médica*

Institución Médica	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Si	0	0	0	0,97	1	18
No	0	0	1	2,06	3	18

Tabla 4.6: *Medidas de resumen de Ccorona según Consume Mate*

Consume Mate	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Si	0	0	1	1,62	2	18
No	0	0	0	0,9	1	11

Tabla 4.7: *Medidas de resumen de Ccorona según Fuma*

Fuma	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Si	0	0	1	2,31	3	18
No	0	0	0	1,16	1	17

Se ajusta un modelo de regresión Binomial Negativo donde las variables significativas al 10 % para modelar Ccorona son Estudio Universitario, Institución Médica

Colectiva, Consume Mate, Fuma e INSE, por lo que se ajusta el modelo nuevamente con estas variables (tabla 4.8).

El modelo ajustado resultante usando la distribución BN es:

$log(Ccorona) = 1{,}019 + 0{,}301 * EstUniv(No) + 0{,}418 * InstMed(No) - 0{,}316 * ConsMate(No) - 0{,}543 * Fuma(No) - 0{,}020 * INSE$

Tabla 4.8: *Modelo Estimado usando Binomial Negativo*

	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	1.019	0.229	4.455	8.37e-06
Estudio Universitario-No	0.301	0.124	2.427	0.015
Institución Médica-No	0.418	0.091	4.611	4.01e-06
Consume Mate-No	-0.316	0.106	-2.997	0.003
Fuma-No	-0.543	0.095	-5.740	9.68e-09
INSE	-0.020	0.004	-5.238	1.62e-07

La estimación de este modelo es similar a los datos reales aunque subestima la cantidad de ceros como se muestra en la figura 4.5.

Por lo expuesto se intentan ajustar modelos para exceso de ceros para mejorar el número estimado de ceros.

En el modelo Cero Inflado Binomial Negativo no existen variables significativas para el componente de cero-inflado, por lo cual se ajusta un modelo Hurdle Binomial Negativo. Se puede ver que las variables significativas al 10 % en el componente de conteo difieren de las significativas para el componente cero.

El modelo ajustado resultante es: Para el componente binario

$log(\frac{\pi}{1-\pi}) = 1{,}092 + 0{,}563 * InstMed(No) - 0{,}317 * ConsMat(No) - 0{,}735 * Fuma(No) - 0{,}022 * INSE$,

para el componente de conteo truncado

$log(Ccorona_t) = 0{,}890 + 0{,}317 * EstUni(No) + 0{,}249 * InstMed(No) - 0{,}399 *$

$Fuma(No) - 0{,}018 * INSE$

Tabla 4.9: *Modelo Estimado Hurdle Binomial Negativa*

Componente Hurdle				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	1.092	0.211	5.171	2.33e-07
Institución Médica-No	0.563	0.118	4.778	1.77e-06
Consume Mate-No	-0.317	0.132	-2.405	0.016
Fuma-No	-0.735	0.130	-5.645	1.66e-08
INSE	-0.022	0.004	-5.369	7.90e-08
Componente Conteo				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	0.890	0.322	2.763	0.006
Estudio Universitario-No	0.317	0.176	1.801	0.072
Institución Médica-No	0.249	0.118	2.103	0.035
Fuma-No	-0.399	0.115	-3.450	0.001
INSE	-0.018	0.006	-3.321	0.001

Comparamos la estimación del Modelo Binomial Negativo con el Modelo Hurdle Binomial Negativa a través de la figura 4.5.

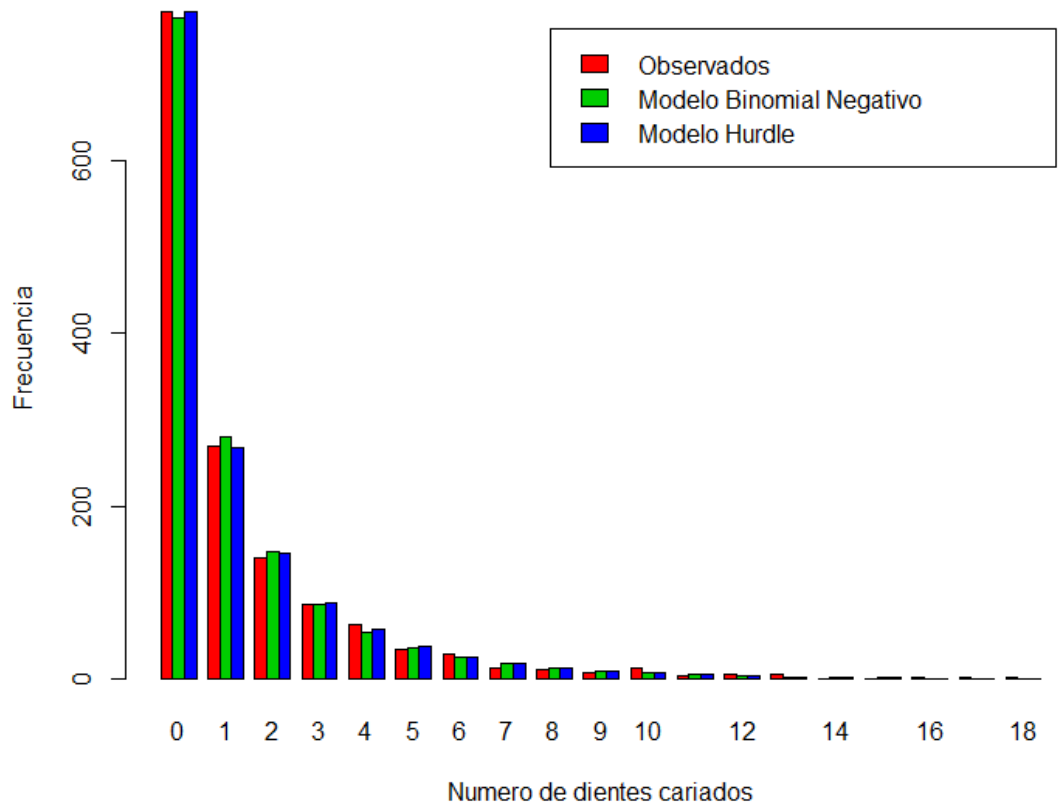

Figura 4.5: *Valores Observados vs. Valores Estimados con Modelos Binomial Negativa y Hurdle Binomial Negativa*

48

4.1.2.1. Evaluación y Validación de los modelos.

Para analizar la validez de los modelos se analizan los residuos hallándose su promedio y su covarianza con los valores ajustados las cuales son $e_{\tilde{BN}}) = -0{,}0009391$ y $Cov(e, \hat{y}) = -0{,}004331$ para el modelo Binomial Negativo y $e_{\tilde{HBN}}) = -0{,}0009362$ y $Cov(e, \hat{y}) = -0{,}004615$ para el modelo Hurdle Binomial Negativo, esto es, muy cercanos a cero.

Para evaluar el ajuste de estos modelos se calculan los errores absolutos con la ecuación (2.1), el cual da un valor de 1.546 para el modelo Binomial Negativo y 1.556 para el modelo Hurdle Binomial Negativo, indicando que sería mejor el ajuste Binomial Negativo; en cambio el test $pseudo-R^2$ da un valor de 0.040 para Binomial Negativo y 0.043 para el modelo Hurdle, y el AIC da 4517 para el modelo Binomial Negativo y 4511 para el modelo Hurdle, sugiriendo lo opuesto, aunque los indicadores para ambos modelos son muy próximos.

4.2. Variable Pcorona (Corona perdida)

Pcorona representa el número de dientes perdidos, o sea el número de piezas dentarias extraídas. Al igual que en la sección anterior se realiza un análisis descriptivo, luego se seleccionan distribuciones que se ajusten a la variable, para posteriormente ajustar modelos de regresión.

Al trabajar con datos sin valores faltantes, para el análisis de la variable Pcorona, se tiene un total de 1350 individuos. Esta variable toma valores de 0 a 32, presentando una distribución bimodal como se muestra en la figura 4.6. Por no ser el objetivo del trabajo analizar el problema de las distribuciones bimodales se elimina el valor 32 de los datos, lo cual desde el punto de vista epidemiológico tiene sentido ya que representa a los individuos edéntulos, por lo cual se trabaja de aquí en adelante con personas que tienen por lo menos una pieza dental.

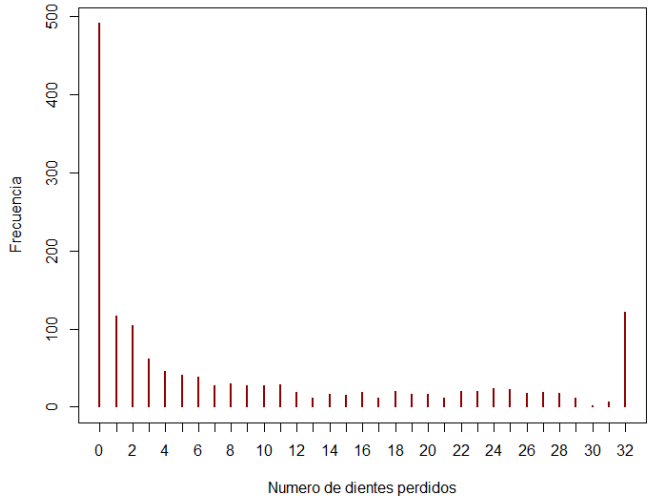

Figura 4.6: *Gráfico de frecuencias absolutas de Pcorona*

Pcorona tiene media 6.74 y varianza 77.28, es decir que su varianza es 11.47 veces su media. Los datos de esta variable se encuentran concentrados en el valor 0 decreciendo hacia el valor 31 como se muestra en la figura 4.7.

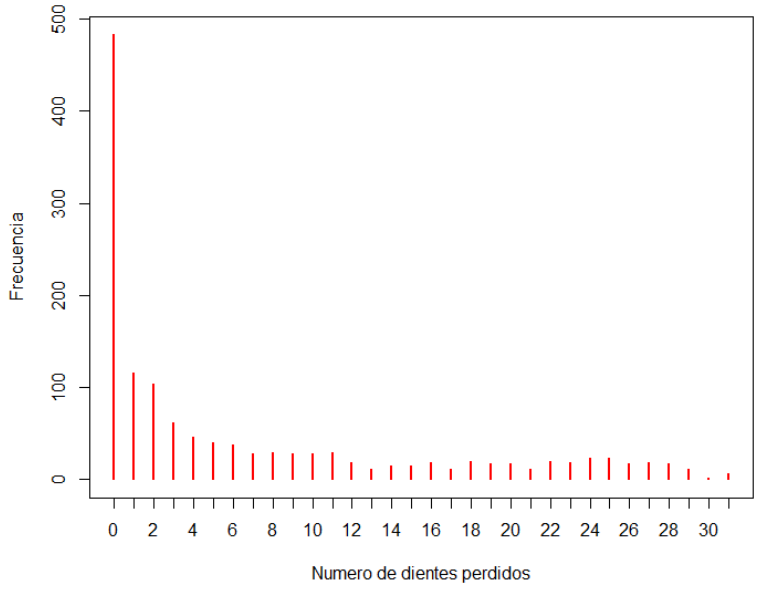

Figura 4.7: *Gráfico de frecuencias de Pcorona*

50

4.2.1. Distribución de Probabilidad para Pcorona

Al tratarse de una variable de conteo, al igual que Ccorona, se comienza ajustando las distribuciones Poisson y Binomial Negativa, cuyos gráficos de ajuste se muestran en la figura 4.8. Como estas tres distribuciones parecen no ajustarse a los datos empíricos y, en particular, subestiman la cantidad de conteos 0, se prueba el ajuste con modelos para exceso de 0's con la distribución de conteo que mejor se ajustó a los datos, Binomial Negativa.

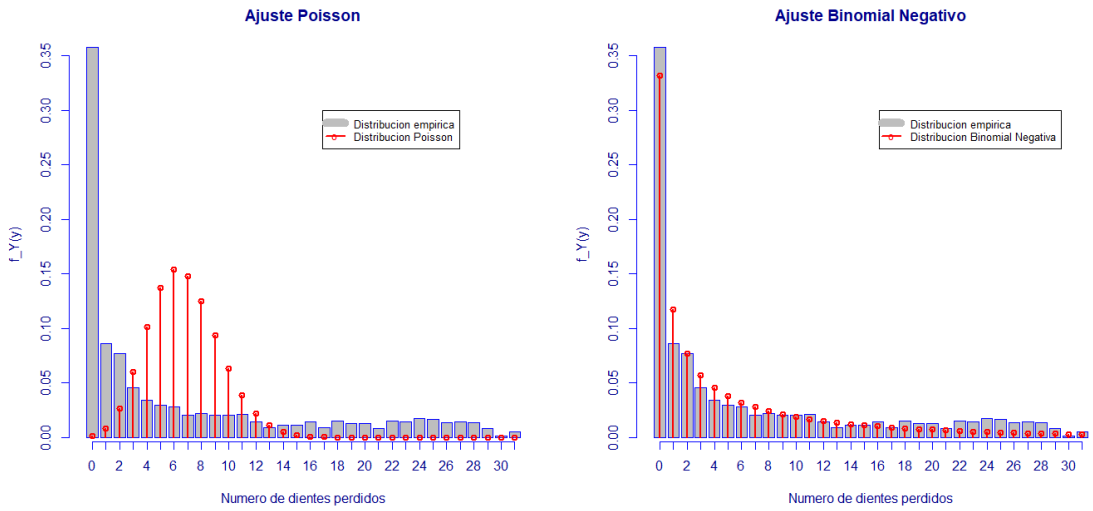

Figura 4.8: *Primeros ajustes a la variable Pcorona*

Así se muestra en las figuras 4.9 y 4.10 el ajuste de los modelos Hurdle Binomial Negativo y Cero Inflado Binomial Negativo.

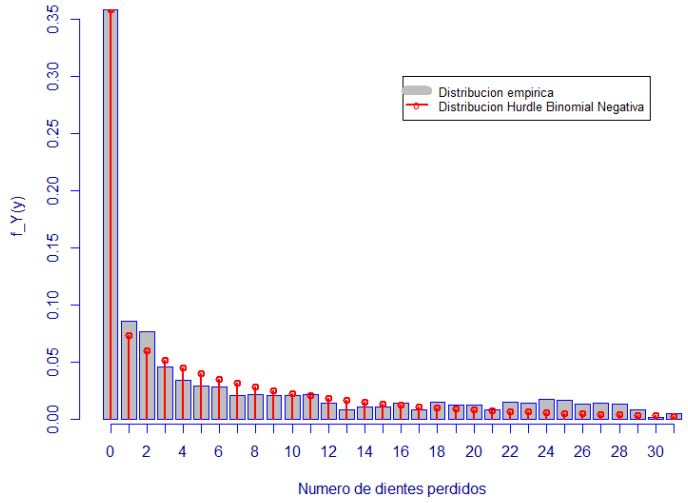

Figura 4.9: *Ajuste Hurdle Binomial Negativo a la variable Pcorona*

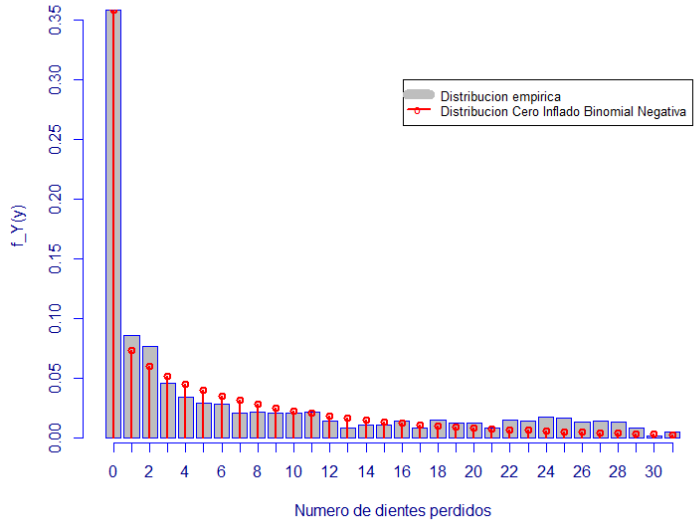

Figura 4.10: *Ajuste Cero Inflado Binomial Negativo a la variable Pcorona*

4.2.2. Modelos de Regresión para Pcorona

Al igual que para Ccorona, se quiere modelar la variable Pcorona a través de modelos de regresión. Se muestra en la figura 4.11 y en las tablas bivariadas las relaciones parciales con las variables explicativas.

Se puede ver que la cantidad de personas sin dientes perdidos disminuye a medida que aumenta la edad, ya que la mitad de los individuos con edades de 15 a 24 años no tienen dientes perdidos, la mitad de los individuos con edades de 35 a 44 años tiene 7 o menos dientes perdidos y la mitad de los individuos con edades de 65 a 74 años tiene 19 o menos dientes perdidos.

El número de personas que no tienen dientes perdidos es mayor si tienen estudios universitarios que si no lo tienen, así como el 75 % de las personas que tienen estudios universitarios tienen 7 o menos dientes perdidos, y el 75 % de las que no tienen estudio universitario tienen 12 o menos dientes perdidos, esto es las personas que no tienen estudios universitarios tienen un número mayor de dientes perdidos.

La cantidad de personas con 2 o menos dientes perdidos no parece diferenciarse según si fuma o no, pero el 75 % de las personas que fuman tienen 9 o menos dientes perdidos y el 75 % de las personas que no fuman tienen 12 o menos dientes perdidos. De igual forma, la cantidad de personas con 2 o menos dientes perdidos no parece diferenciarse según el sexo, aunque el 75 % de las personas del sexo femenino tienen 13 o menos dientes perdidos y el 75 % de las personas de sexo maculino tienen 9 o menos dientes perdidos. También sigue este comportamiento la variable institución médica colectiva, ya que la cantidad de personas con 2 o menos dientes perdidos representa el 50 % de las personas tanto con institución médica colectiva como sin la misma, y el 75 % de las personas con institución médica colectiva tienen 10 o menos dientes perdidos y el 75 % de las personas que no tienen institución médica colectiva tienen 12 o menos dientes perdidos.

El 50 % de las personas de Montevideo tienen 2 o menos dientes perdidos y el 50 % de las personas del interior tienen 3 o menos dientes perdidos. El 50 % de las personas que consumen mate tienen 3 o menos dientes perdidos y el 50 % de las personas que no consumen mate no tienen dientes perdidos, o sea que las personas que consumen mate tienen más dientes perdidos.

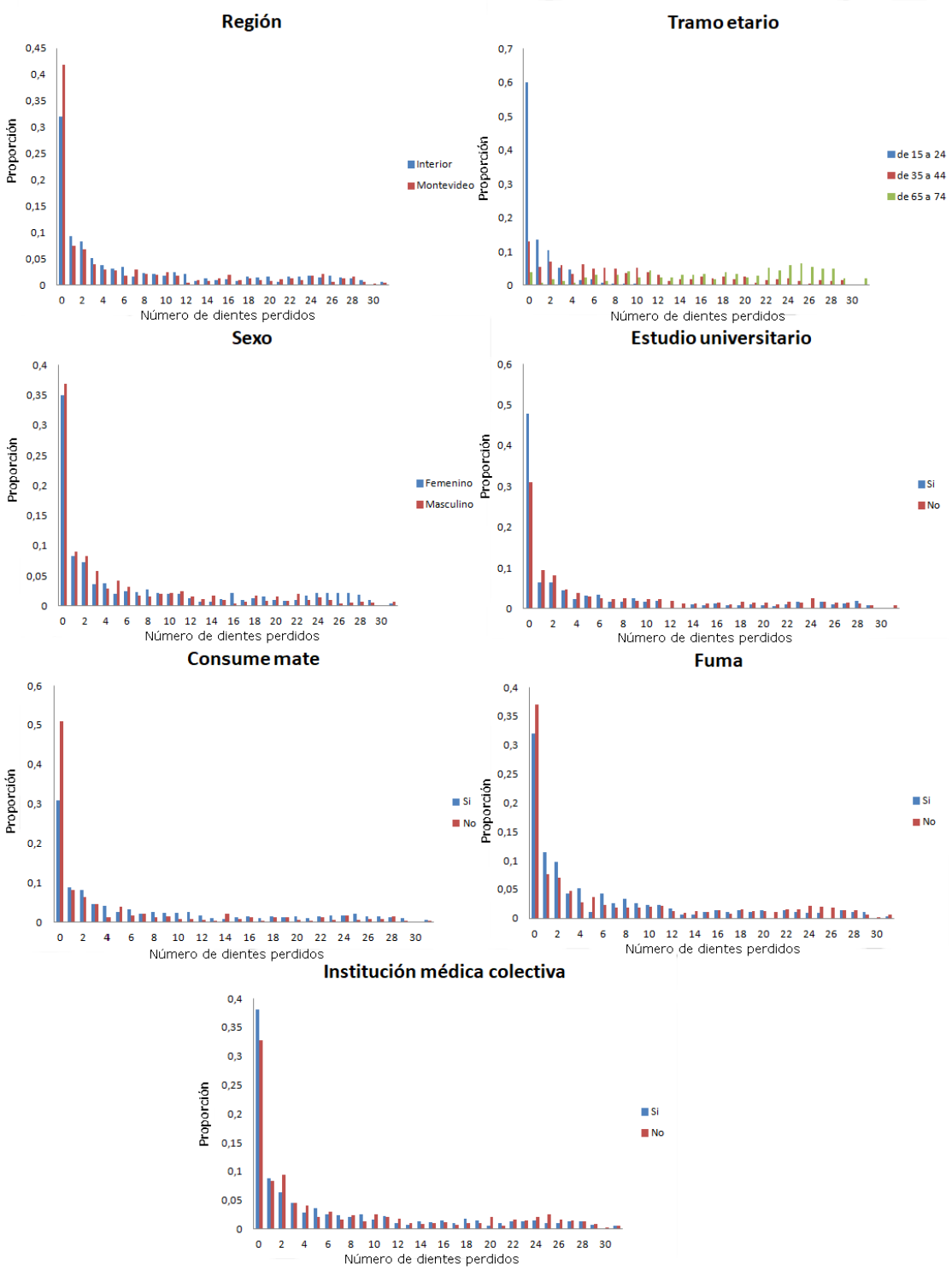

Figura 4.11

54

Tabla 4.10: *Medidas de resumen de Pcorona según Región*

Región	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Montevideo	0	0	2	6,21	10	31
Interior	0	0	3	7,07	11	31

Tabla 4.11: *Medidas de resumen de Pcorona según Tramo Etario*

Tramo Etario	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
15-24	0	0	0	1,17	2	20
35-44	0	2	7	9,28	15	31
65-74	0	11	19	17,33	25	31

Tabla 4.12: *Medidas de resumen de Pcorona según Sexo*

Sexo	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Femenino	0	0	2	7,46	13	31
Masculino	0	0	2	5,80	9	31

Tabla 4.13: *Medidas de resumen de Pcorona según Estudio Universitario*

Estudio Universitario	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Si	0	0	1	5,22	7	29
No	0	0	3	7,35	12	31

Tabla 4.14: *Medidas de resumen de Pcorona según Institución Médica*

Institución Médica	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Si	0	0	2	6,37	10	31
No	0	0	2	7,23	12	31

Tabla 4.15: *Medidas de resumen de Pcorona según Consume Mate*

Consume Mate	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Si	0	0	3	7,40	12	31
No	0	0	0	4,72	5,75	31

Tabla 4.16: *Medidas de resumen de Pcorona según Fuma*

Fuma	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Si	0	0	2	5,96	9	31
No	0	0	2	7,01	12	31

Tomando en cuenta los ajustes de las distribuciones seleccionadas anteriormente por su ajuste a Pcorona, se ajustan los modelos de regresión Cero Inflado Binomial Negativo y Hurdle Binomial Negativo.

Tabla 4.17: *Modelo Estimado Cero Inflado Binomial Negativa Pcorona*

Componente Cero Inflado				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	-0.933	0.427	-2.184	0.029
Región-Montevideo	0.396	0.177	2.241	0.025
Tramo Etario-de 35 a 44	-2.234	0.220	-10.16	$<$ 2e-16
Tramo Etario-de 65 a 74	-3.821	0.374	-10.21	$<$ 2e-16
Estudio Universitario-No	-0.787	0.231	-3.410	0.001
Consume Mate-No	0.684	0.189	3.612	3.04e-4
Fuma-No	0.619	0.208	2.978	0.003
INSE	0.021	0.007	2.777	0.005
Componente de Conteo				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	1.350	0.110	12.29	$<$ 2e-16
Tramo Etario-de 35 a 44	1.520	0.068	22.40	$<$ 2e-16
Tramo Etario-de 65 a 74	2.133	0.071	29.87	$<$ 2e-16
Sexo-M	-0.165	0.050	-3.327	0.001
Institución Médica-No	0.175	0.053	3.299	0.001
Fuma-No	-0.148	0.059	-2.487	0.013
INSE	-0.013	0.002	-6.349	2.16e-10

Las variables significativas al 10 % para el modelo Cero Inflado Binomial Negativo son tramo etario, sexo, institución médica colectiva, fuma e INSE para el componente de conteo, y región, tramo etario, estudio universitario, consume mate, fuma e INSE para el componente binario.

Se observa que las variables significativas no son las mismas para el componente de conteo que para el componente cero inflado. Se realiza una nueva estimación con las variables significativas (tabla 4.17).

Con este último modelo se predicen los valores de Pcorona los cuales se muestran en la figura 4.12.

Luego se ajusta un modelo Hurdle Binomial Negativo, siendo las variable significativas al 10 % tramo etario, sexo, institución médica colectiva, fuma e INSE para el componente de conteo y región, tramo etario, estudio universitario, consume mate, fuma e INSE para el componente cero inflado.

Las variables significativas son las mismas que en el modelo Cero Inflado Binomial Negativo en ambos componentes. Se estima un modelo Hurdle con estas variables que

se muestra en la tabla 4.18, luego se predicen los valores para Pcorona y se comparan en la tabla 4.12. Los dos modelos predicen valores muy similares y cercanos a los empíricos.

Tabla 4.18: *Modelo Estimado Hurdle Binomial Negativa Pcorona*

Componente Hurdle				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	0.632	0.360	1.756	0.079
Región-Montevideo	-0.340	0.151	-2.257	0.024
Tramo Etario-de 35 a 44	2.526	0.194	13.02	$<$ 2e-16
Tramo Etario-de 65 a 74	4.083	0.333	12.24	$<$ 2e-16
Estudio Universitario-No	0.679	0.203	3.343	0.001
Consume Mate-No	-0.604	0.165	-3.652	2.60e-04
Fuma-No	-0.575	0.164	-3.502	4.61e-04
INSE	-0.024	0.006	-3.825	1.31e-04
Componente de Conteo				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	1.363	0.110	12.44	$<$ 2e-16
Tramo Etario-de 35 a 44	1.518	0.068	22.39	$<$ 2e-16
Tramo Etario-de 65 a 74	2.132	0.071	29.88	$<$ 2e-16
Sexo-M	-0.173	0.050	-3.433	0.001
Institución Médica-No	0.175	0.054	3.256	0.001
Fuma-No	-0.149	0.059	-2.510	0.012
INSE	-0.013	0.002	-6.496	8.26e-11

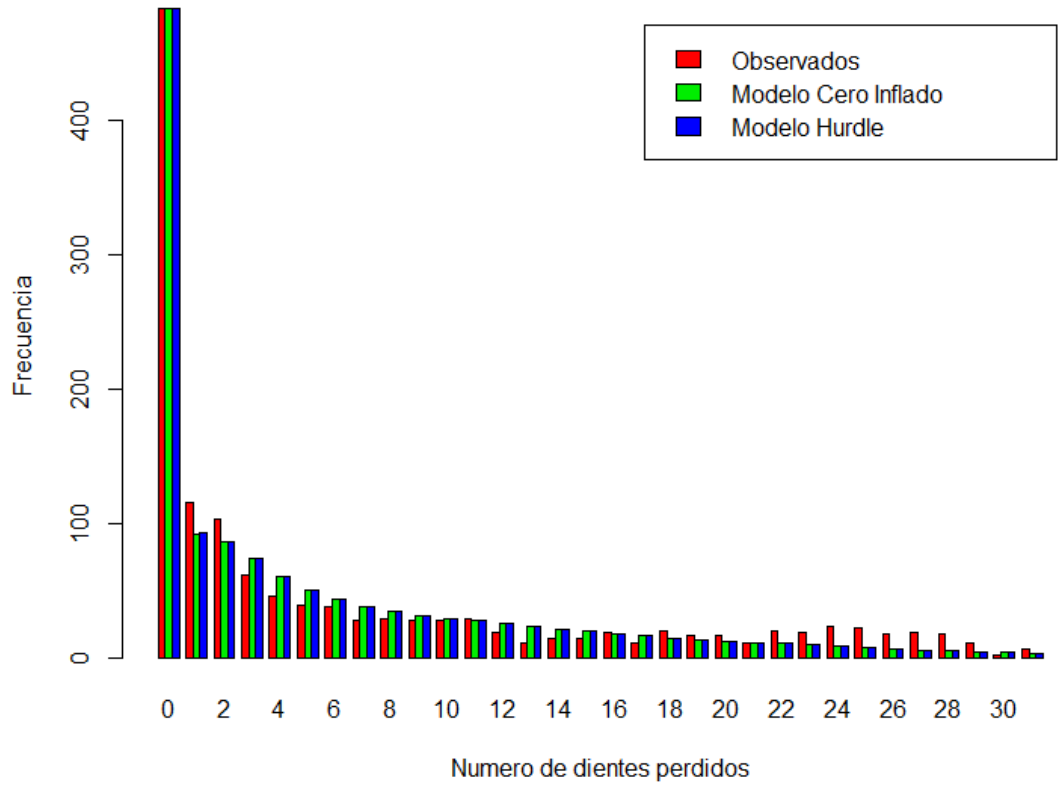

Figura 4.12: *Valores Observados vs. Valores Estimados Cero Inflado y Valores Estimados Hurdle*

4.2.2.1. Evaluación y Validación del modelo

Se analiza la validez del modelo a través de los promedios y varianzas de sus errores, las cuales son para el modelo Cero Inflado Binomial Negativo $e\tilde{}_{ZI} = 0{,}036$ y $Cov(e, \hat{y}) = -2{,}29$ y para el modelo Hurdle Binomial Negativo $e\tilde{}_{H} = 0{,}037$ y $Cov(e_H, \hat{y}_H) = -2{,}45$.

Luego, para evaluar la calidad del ajuste de estos modelos se calculan los errores con la ecuación (2.1), el cual es 3.613 para el modelo Cero Inflado Binomial Negativo y 3.634 para el modelo Hurdle Binomial Negativo, y se realiza el test $pseudo - R^2$ que da un valor de 0.16736 para el modelo Cero Inflado Binomial Negativo y 0.16733 para el modelo Hurdle Binomial Negativo, así como el AIC es 6258.69 para el modelo Cero Inflado y 6258.86 para el modelo Hurdle. En los tres casos el modelo Cero Inflado Binomial Negativo muestra mejores resultados, aunque los indicadores son muy similares para ambos modelos.

4.3. Variable Ocorona (Corona obturada)

Ocorona representa el número de dientes obturados, es decir, el número de dientes que fueron tratados por caries dental. Luego de un análisis descriptivo se seleccionan las distribuciones que mejor se ajusten a la variable para, con éstas, ajustar modelos de regresión.

Se trabaja con datos sin valores faltantes, por lo que para la variable Ocorona se tiene un total de 1469 individuos. Esta variable toma valores de 0 a 31, tiene media 2.39 y varianza 12.63, por lo que su varianza es 5.28 veces su media. Los datos se encuentran concentrados en el valor cero, decreciendo hacia el valor máximo, 31, como se puede ver en la figura 4.13.

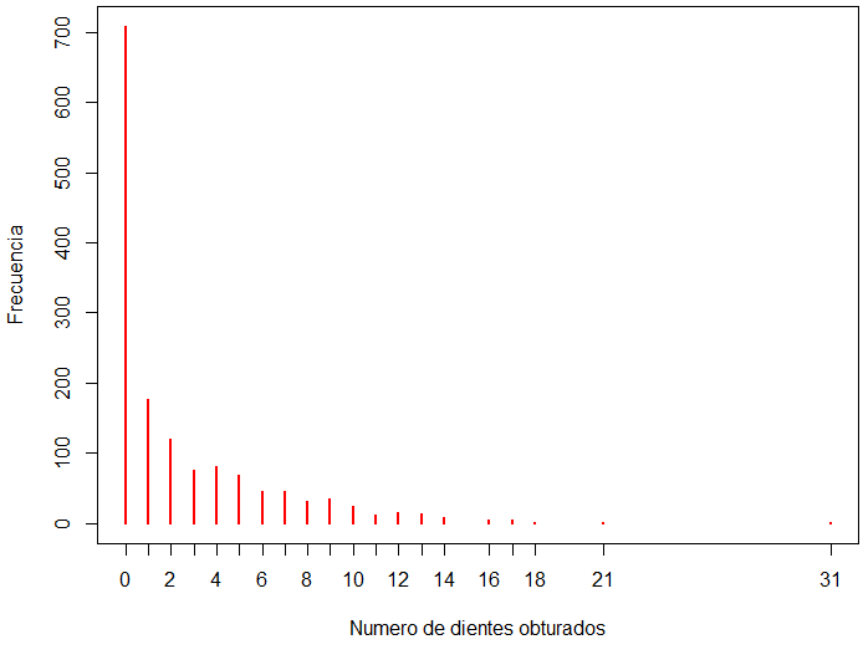

Figura 4.13: *Gráfico de frecuencias de Ocorona*

4.3.1. Distribución de Probabilidad para Ocorona

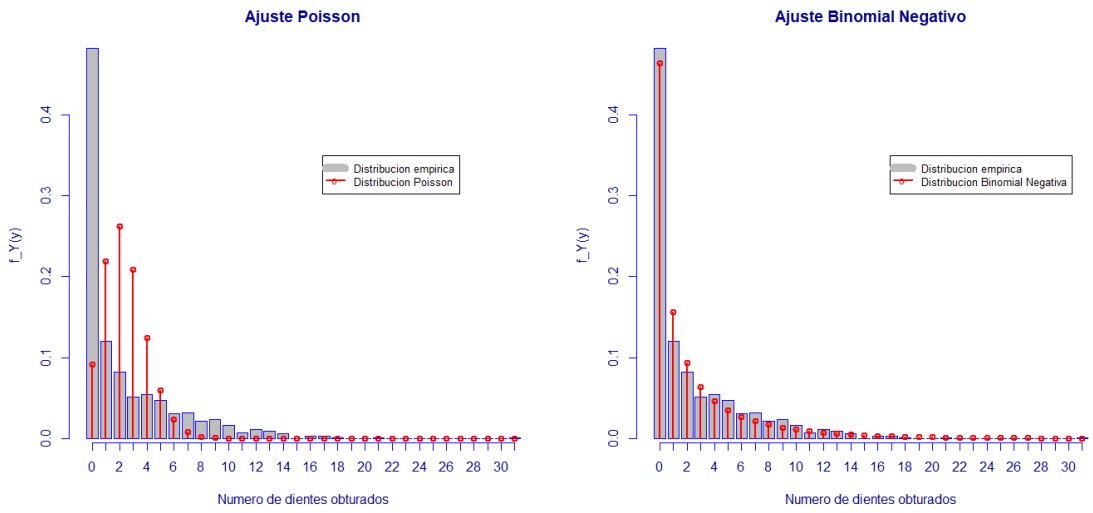

Figura 4.14: *Primeros Ajustes a la variable Ocorona*

Al igual que las variables a explicar anteriores, ésta también es una variable de conteo por lo que se intentan ajustar las distribuciones Poisson y Binomial Negativa, como se muestra en el gráfico 4.14. Se puede ver que estas dos distribuciones no se ajustan correctamente a la variable Ocorona, por lo que se ajustan modelos de dos componentes.

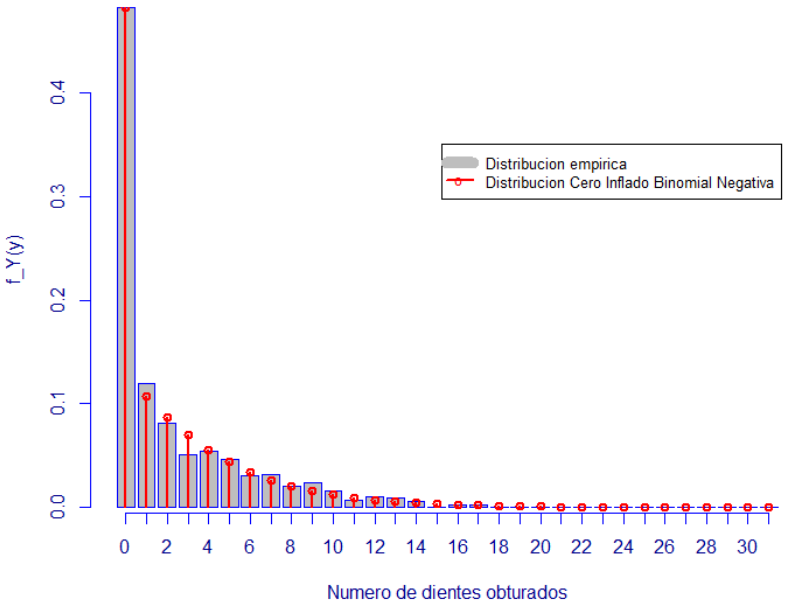

Figura 4.15: *Ajuste Cero Inflado Binomial Negativo a la variable Ocorona*

60

En las figuras 4.15 y 4.16 se muestran los ajustes de los modelos Cero Inflado Binomial Negativo y Hurdle Binomial Negativo.

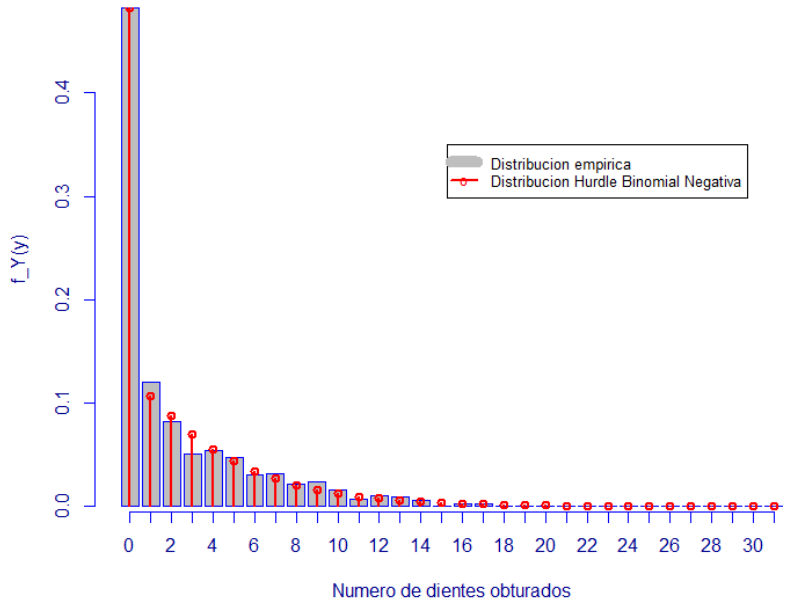

Figura 4.16: *Ajuste Hurdle Binomial Negativo a la variable Ocorona*

4.3.2. Modelos de Regresión para Ocorona

Para modelar la variable Ocorona a través de modelos de regresión se trabaja con los modelos Hurdle Binomial Negativo y Cero Inflado Binomial Negativo que fueron los que mejor se ajustaron a los datos empíricos.

En la figura 4.17 y en las tablas bivariadas se muestran las relaciones parciales de la variable Ocorona con las variables explicativas.

Se puede notar que el número de personas que no tienen dientes obturados es mayor si no tienen estudios universitarios que si lo tienen, así como el 75 % de las personas que no tienen estudios universitarios tienen 3 o menos dientes obturados, y el 75 % de las que tienen estudios universitarios tienen 6 o menos dientes obturados, esto es las personas que no tienen estudios universitarios tienen un número menor de dientes obturados. Lo mismo sucede con las personas según si tienen institución médica colectiva, el 75 % de las personas que tienen institución médica colectiva tienen 5 o menos dientes obturados, y el 75 % de las que no tienen institución médica colectiva tienen 2 o menos dientes obturados.

El número de personas que no tienen dientes obturados es mayor en el grupo de las que fuman que dentro de las que no fuman, el 75 % de las personas que fuman tienen 2 o menos dientes obturados y el 75 % de las personas que no fuman tienen 4 o menos dientes obturados, o sea que las personas que no fuman tienen más dientes obturados. El 75 % de las personas de sexo masculino tienen 3 o menos dientes obturados y el 75 % de las personas de sexo femenino tienen 4 o menos dientes obturados. La cantidad de personas con dientes obturados o sin éstos no parece diferenciarse según los tramos de edad de 15 a 24 y de 65 a 74, aunque si pertenecen al tramo de 35 a 44 se puede ver una menor cantidad de individuos sin dientes obturados.

La cantidad de personas con dientes obturados o sin éstos no parece diferenciarse según si consume mate o no, ya que tanto el primer cuartil, la mediana y el tercer cuartil de ambos grupos coinciden.

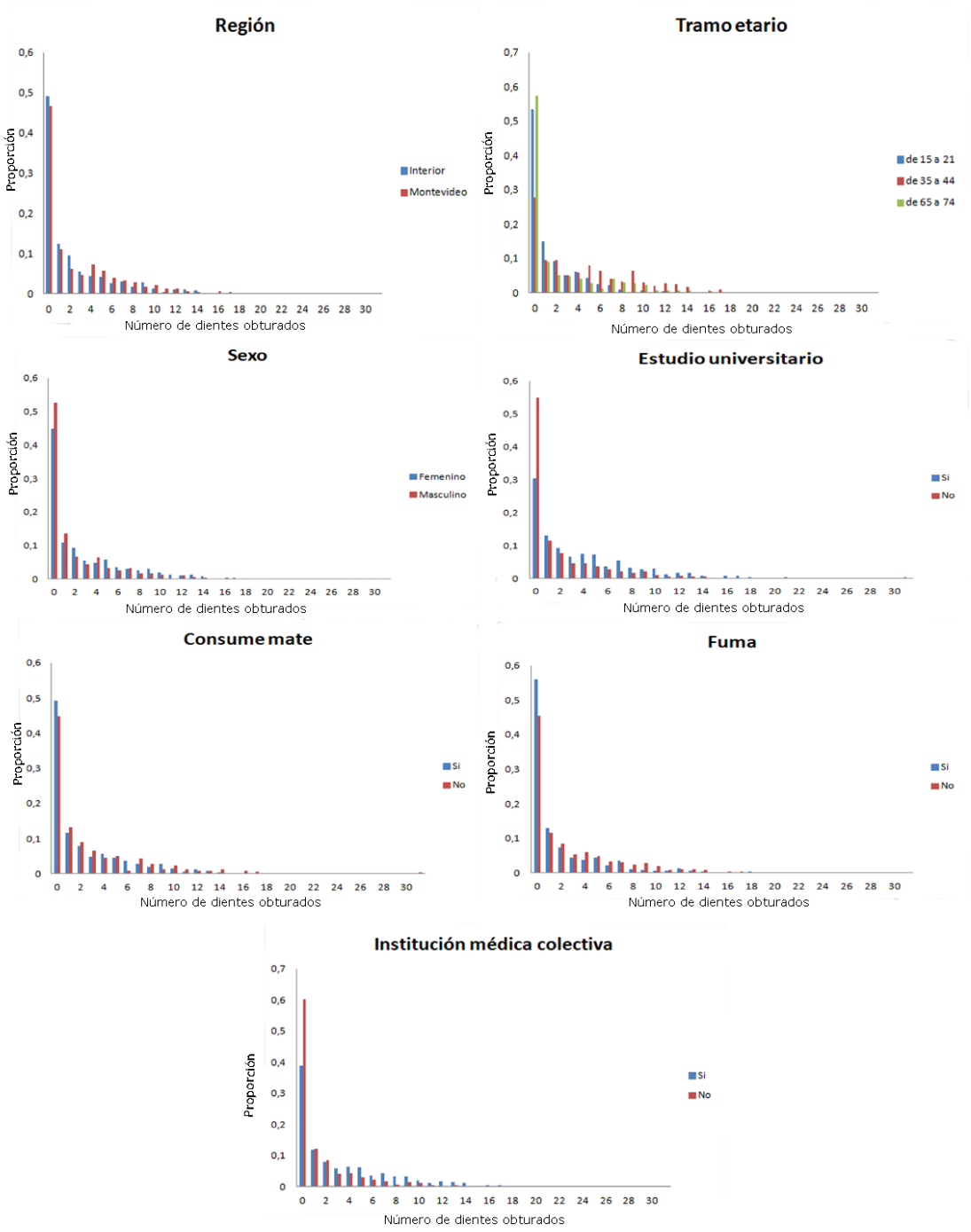
Región
Proporción
Número de dientes obturados
Interior
Montevideo
Tramo etario
de 15 a 21
de 35 a 44
de 65 a 74
Sexo
Femenino
Masculino
Estudio universitario
Si
No
Consume mate
Fuma
Institución médica colectiva

Figura 4.17

Tabla 4.19: *Medidas de resumen de Ocorona según Región*

Región	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Montevideo	0	0	1	2,61	4	18
Interior	0	0	1	2,26	3	31

Tabla 4.20: *Medidas de resumen de Ocorona según Tramo Etario*

Tramo Etario	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
15-24	0	0	0	1,48	2	13
35-44	0	0	3	4,35	7	21
65-74	0	0	0	2,26	3	31

Tabla 4.21: *Medidas de resumen de Ocorona según Sexo*

Sexo	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Femenino	0	0	1	2,69	4	21
Masculino	0	0	0	1,99	3	31

Tabla 4.22: *Medidas de resumen de Ocorona según Estudio Universitario*

Estudio Universitario	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Si	0	0	2	3,70	6	31
No	0	0	0	1,90	3	17

Tabla 4.23: *Medidas de resumen de Ocorona según Institución Médica*

Institución Médica	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Si	0	0	1	3,18	5	31
No	0	0	0	1,38	2	13

Tabla 4.24: *Medidas de resumen de Ocorona según Consume Mate*

Consume Mate	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Si	0	0	1	2,31	4	21
No	0	0	1	2,65	4	31

Tabla 4.25: *Medidas de resumen de Ocorona según Fuma*

Fuma	Mín.	Pr. Cuart.	Mediana	Media	Ter. Cuart.	Máx.
Si	0	0	0	1,75	2	18
No	0	0	1	2,61	4	31

Al ajustar el modelo Cero Inflado Binomial Negativo las variables significativas al 10 % fueron Tramo Etario, Sexo, Estudio Universitario, Institución Médica Colectiva, Fuma e INSE para el componente cero inflado y Tramo Etario, Sexo, Institución Médica Colectiva e INSE para el componente de conteo.

Se observa que las variables significativas para el componente binario no son las mismas que para el componente de conteo. Con estas variables se realiza una nueva estimación que se puede ver en la tabla 4.26.

Con esta estimación se predicen los valores de Ocorona cuyo resultado se muestra en la figura 4.18.

Tabla 4.26: *Modelo Estimado Cero Inflado Binomial Negativa*

Componente Cero Inflado				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	0.037	0.430	0.088	0.929
Tramo Etario-de 35 a 44	-0.869	0.205	-4.238	2.26e-05
Tramo Etario-de 65 a 74	0.687	0.192	3.573	3.54e-04
Sexo-M	0.278	0.154	1.798	0.072
Estudio Universitario-No	0.646	0.217	2.968	0.002
Institución Médica-No	0.583	0.163	3.568	3.60e-04
Fuma-No	-0.541	0.173	-3.121	0.001
INSE	-0.028	0.007	-3.711	2.07e-04
Componente de Conteo				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	0.641	0.128	4.973	6.59e-07
Tramo Etario-de 35 a 44	0.758	0.081	9.309	$<$ 2e-16
Tramo Etario-de 65 a 74	0.594	0.091	6.501	8.00e-11
Sexo-M	-0.179	0.068	-2.610	0.009
Institución Médica-No	-0.236	0.080	-2.943	0.003
INSE	0.010	0.002	4.565	5.01e-06

Luego se ajusta un modelo Hurdle Binomial Negativo, en el cual las variables significativas al 10 % son: para el componente binario tramo etario, sexo, estudio universitario, institución médica colectiva, fuma e INSE, y para el componente de conteo tramo etario, sexo, institución médica colectiva e INSE.

Las variables significativas no son las mismas para el componente binario que para el componente de conteo, pero son las mismas que en el modelo Cero Inflado. Con estas variables se realiza una nueva estimación (tabla 4.27) con la cual se predicen los valores de Ocorona y se comparan con los empíricos.

Tabla 4.27: *Modelo Estimado Hurdle Binomial Negativa*

Componente Hurdle				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	-0.467	0.328	-1.424	0.154
Tramo Etario-de 35 a 44	1.133	0.149	7.625	2.44e-14
Tramo Etario-de 65 a 74	-0.253	0.137	-1.842	0.065
Sexo-M	-0.334	0.117	-2.857	0.004
Estudio Universitario-No	-0.512	0.161	-3.187	0.001
Institución Médica-No	-0.570	0.123	-4.622	3.80e-06
Fuma-No	0.405	0.136	2.974	0.002
INSE	0.023	0.005	4.276	1.90e-05
Componente de Conteo				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	0.628	0.128	4.894	9.88e-07
Tramo Etario-de 35 a 44	0.716	0.077	9.257	$<$ 2e-16
Tramo Etario-de 65 a 74	0.560	0.088	6.370	1.88e-10
Sexo-M	-0.164	0.069	-2.385	0.017
Institución Médica-No	-0.254	0.079	-3.219	0.001
INSE	0.012	0.002	5.090	3.59e-07

Los dos modelos predicen valores muy similares y cercanos a los empíricos como se puede ver en la figura 4.18.

66

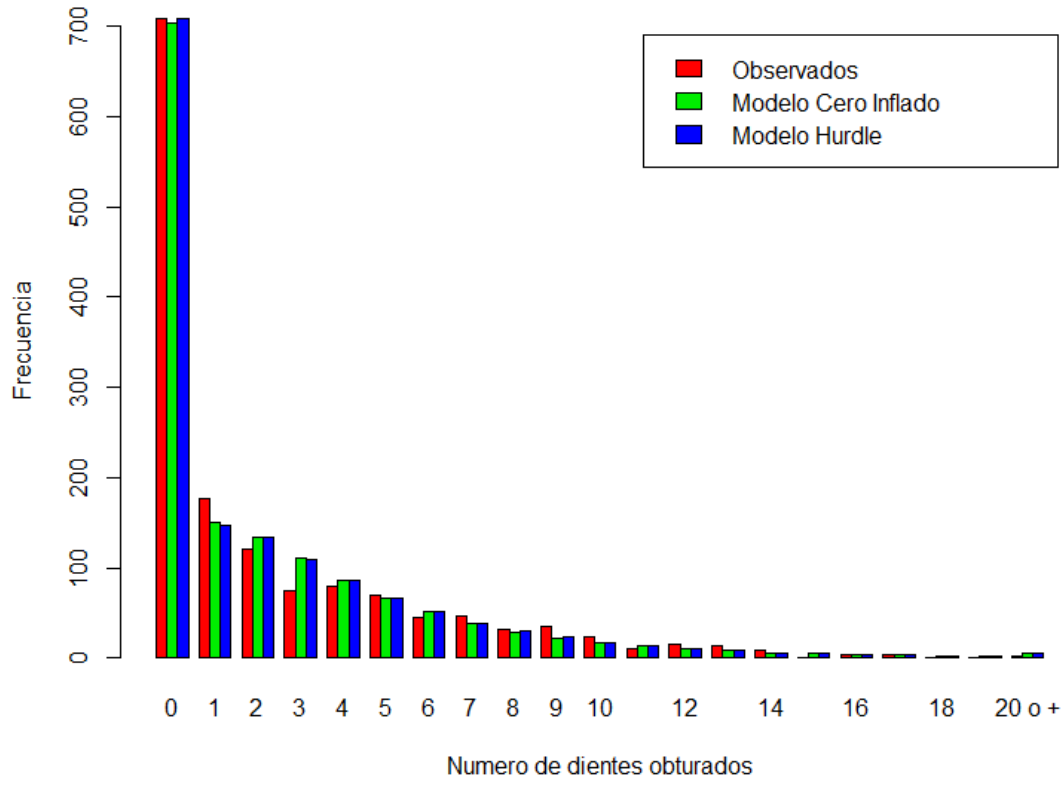

Figura 4.18: *Valores Observados vs. Valores Estimados Cero Inflado y Valores Estimados Hurdle*

Las diferencias en las cantidades totales estimadas con las reales se debe a que son la suma de probabilidades de cada individuo de tener cierta cantidad de dientes obturados.

4.3.2.1. Evaluación y Validación del modelo

A través de los residuos se analiza la validez del modelo, viendo sus medidas de resumen. Estos son $e\tilde{}_{ZI} = 0{,}0206$ y $Cov(e, \hat{y}) = 0{,}31$ y para el modelo Hurdle Binomial Negativo $e\tilde{}_H = 0{,}0210$ y $Cov(e_H, \hat{y}_H) = 0{,}28$.

Para evaluar la calidad de ajuste de los modelos se calculan los errores a través de la ecuación (2.1), el cual da un valor de 2.20 para el modelo Cero Inflado Binomial Negativo y 2.21 para el modelo Hurdle Binomial Negativo, el test $pseudo - R^2$ da un valor de 0.0691 para el modelo Cero Inflado y 0.0687 para el modelo Hurdle y el AIC da un valor de 5340 para el modelo Cero Inflado y 5342 para el modelo Hurdle. En

los tres casos el modelo Cero Inflado Binomial Negativo muestra mejores resultados, aunque los indicadores son muy similares para ambos modelos.

4.4. CPOcorona

La variable CPOcorona es el resultado de la suma de las tres variables vistas anteriormente. Como se puede ver en su histograma en la figura 4.19, esta variable es bimodal y no tiene un comportamiento que se asemeje a ninguna distribución conocida con las cuales se trabaja en este informe. Por este motivo no se analiza esta variable en este trabajo, sólo se presenta a efectos informativos.

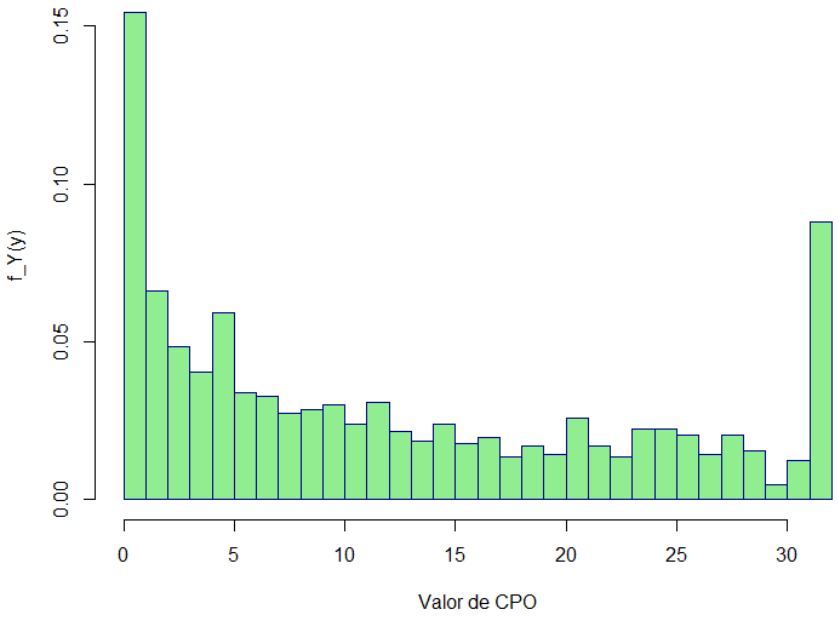

Figura 4.19: *Histograma de CPOcorona*

4.5. Resumen

- Para la variable Ccorona se selecciona el modelo de regresión Binomial Negativo por el principio de parsimonia ya que los indicadores de bondad de ajuste entre éste y el MH Binomial Negativo son muy similares al igual que las estimaciones con ambos modelos.

  El modelo seleccionado es:

  $log(Ccorona) = 1{,}019 + 0{,}301 * EstUniv(No) + 0{,}418 * InstMed(No) - 0{,}316 * ConsMate(No) - 0{,}543 * Fuma(No) - 0{,}020 * INSE$

  donde Ccorona tiene distribución Binomial Negativa.

  Para este modelo, el logaritmo del número de caries aumenta si el encuestado no tiene estudios universitarios, manteniendo las demás variables constantes. Lo mismo sucede cuando no cuenta con institución médica. Por el contrario, el logaritmo del valor esperado del número de caries disminuye a mayores valores de INSE, manteniendo las demás variables constantes, lo mismo sucede cuando el individuo no fuma o no consume mate.

- Para la variable Pcorona se seleccionan los modelos de regresión Hurdle Binomial Negativo y Cero Inflado Binomial Negativo. Dado que los indicadores de bondad de ajuste de éstos son muy similares, al igual que las estimaciones, no es posible seleccionar entre uno de ellos. Además para ambos modelos las variables significativas son las mismas tanto para el componente de conteo como para el componente binario.

  Para estos modelos se concluye que a mayor tramo etario, el peso del componente binario para los dos modelos disminuye. Lo mismo ocurre si el individuo no tiene estudios universitarios, fuma o toma mate. Por el contrario esta probabilidad aumenta a mayores valores de INSE y si el encuestado pertenece a Montevideo.

  Por medio del componente de conteo se observa también que a medida que cambia el tramo etario, el logaritmo del valor esperado de dientes perdidos aumenta, de la misma manera que se concluyó para el componente binario. Este valor también aumenta si el encuestado es de sexo femenino, o no cuenta con institución médica o fuma, así como disminuye a mayores valores de INSE.

- Con respecto a la variable Ocorona, ocurre lo mismo que con la variable Pcorona: no es posible seleccionar entre los modelos Hurdle Binomial Negativo y Cero Inflado Binomial Negativo dada la similitud en los indicadores de bondad

de ajuste y en las variables que resultaron significativas.

En el componente binario, a diferencia de la variable anterior, si el individuo se encuentra en el tramo etario medio (35 a 44), la probabilidad de ningún diente obturado disminuye, pero aumenta si se encuentra en el tramo etario de 65 a 74. Esto se puede deber a que el componente perdido para ese sector de la población es muy alto. Lo mismo sucede con los individuos de sexo masculino, o con los que no cuentan con estudios universitarios o institución médica, a la vez que aumenta para aquellos que fuman y aumenta a mayores valores de INSE.

Por el contrario, el logaritmo del valor esperado de dientes obturados aumenta para los individuos de sexo femenino y para los que cuentan con institución médica, y aumenta también para mayores valores de INSE.

72

Capítulo 5

Conclusiones

En la realización de este trabajo se lograron construir modelos para explicar los componentes del Índice CPO a partir de una muestra que consta de 1485 datos relevados, y tomados de 15 ciudades: Montevideo y 14 ciudades del interior que tienen más de 20.000 habitantes. Se intentó que dichos modelos, para cada uno de los componentes, fuese el que mejor se adaptase a los datos y mejor predijese futuras observaciones, a la vez de cumplir con el principio de parsimonia. Los tres modelos ajustados son del tipo mixto, lo que es de gran ayuda para hacer frente a la sobredispersión que se presenta por lo general al trabajar con datos reales.

5.1. Conclusiones para Ccorona

Se encontró que con un modelo de regresión Binomial Negativo se puede explicar la variable Ccorona a través de las variables explicatvas Estudio Universitario, Institución Médica, Consume Mate, Fuma e INSE, las que son significativas al 10 %.

Para este modelo, un coeficiente positivo aumenta el logaritmo del número de caries, por lo que si el encuestado no tiene estudios universitarios, no cuenta con institución médica, fuma o consume mate, este logaritmo aumenta. Por el contrario, a mayores valores de INSE, este logaritmo disminuye.

5.2. Conclusiones para Pcorona

Para esta variable se encontró que con modelos Cero Inflado Binomial Negativo y Hurdle Binomial Negativo se logra el mejor ajuste de la distribución a los datos,

siendo ambos ajustes muy similares. Además también se observa que las variables que explican los 2 modelos en sus 2 componentes son las mismas.

Las variables que resultaron significativas para el componente binario para esta variable son tramo etario, estudios universitarios, fuma, toma mate, región e INSE.

Además, para el componente de conteo las variables signifiativas son tramo etario, sexo, institución médica, fuma e INSE.

5.3. Conclusiones para Ocorona

Al igual que para la variable Pcorona, para Ocorona se considera el ajuste con las distribuciones Hurdle BN y Cero Inflado BN. Ambos modelos producen ajustes muy similares. Además las variables que explican los 2 modelos en sus 2 componentes son las mismas.

En el componente binario las variables significativas son tramo etario, sexo, estudios universitarios, institución médica, fuma e INSE.

Para el componente binario las variables significativas son tramo etario, sexo, institución médica e INSE.

5.4. Conclusiones generales

Se encontró que las variables que explican el comportamiento de uno de los componentes del índice CPO no son las mismas que las que explican el comportamiento de los otros.

Consideraciones a Futuro

Como futuros pasos se propone:

- analizar si existe sobredispersión en los datos trabajados o hay lo que se conoce como "sobredispersión aparente"(Hilbe, 2014).
- evaluar si existen otras distribuciones de tipo discretas que puedan ajustar adecuadamente estas variables de conteo, en especial la variable CPO, la que no presenta un patrón de que se asemeje a alguna distribución conocida; y P, que presenta comportamiento bimodal.
- evaluar la calidad de predicción de futuros por medio de muestras de validación.

- considerar el diseño muestral como parte del análisis.

76

Bibliografía

[1] (2008). *Programa Nacional de Salud Bucal*.

[2] A. Colin Cameron, P. K. T. (2013). *Regression Analysis of Count Data*. Cambridge University Press.

[3] Chaves, M. M. (1962). Odontología sanitaria. *Publicaciones Científicas*, (63).

[4] Fernández, A. & Perera, M. (2003). Indice de niveles socieoconómicos (inse). Technical report, CPA/FERRERE.

[5] GRUEBBEL, A. O. (1944). A measurement of dental caries prevalence and treatment service for deciduous teeth. *Journal of Dental Research*, Vol.23:pp.163–168.

[6] Hilbe, J. (2011). *Negative binomial regression*. Cambridge University Press, Cambridge, UK New York.

[7] Hilbe, J. M. (2014). *Modeling Count Data*. Cambridge University Press.

[8] Jackman, S. (2015). *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University*. Department of Political Science, Stanford University, Stanford, California. R package version 1.4.9.

[9] Klein, H., Palmer, C. E., and Knutson, J. W. (1938). Studies on dental caries: I. dental status and dental needs of elementary school children. *Public Health Reports (1896-1970)*, 53(19):751.

[10] Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1.

[11] Leiva, V., Hernández, H., and Sanhueza, A. (2008). An R Package for a general class of inverse gaussian distributions. *Journal of Statistical Software*, 26(4).

[12] Lorenzo, S., Álvarez Vaz, R., and Blanco, S. amd Pérez, M. (2013). Primer re-

levamiento nacional de salud bucal en población joven y adulta uruguaya. *Odontoestomatología*, 15.

[M. Fernández PratTs] M. Fernández PratTs, M. Barciela González-Longoria, e. a. Indices epidemiológicos para medir la caries dental. Technical report, Benemérita Universidad Autónoma de Puebla, Facultad de Estomatología.

[14] Mangiafico, S. (2017). *rcompanion: Functions to Support Extension Education Program Evaluation*. R package version 1.10.1.

[15] Meyer, D., Zeileis, A., and Hornik, K. (2016). *vcd: Visualizing Categorical Data*. R package version 1.4-3.

[16] Moscote, O. y Arley, W. (2012). Modelo logit y probit: un caso de aplicación. *Comunicaciones en Estadística*, 5(2):123–134.

[17] Organización Mundial de la Salud, G. ., editor (1997). *Encuestas de Salud Bucal*.

[18] P. Olmos, S. Piovesan, e. a. (2013). Caries dental. la enfermedad oral más prevalente: Primer estudio poblacional en jóvenes y adultos uruguayos del interior del país. *Odontoestomatología*, 15.

[19] R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[20] Ricci, V. (2005). Fitting distributions with r. *R Project*.

[21] Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554.

[22] Shaban, S. A. (1981). On the discrete poisson-inverse gaussian distribution. *Biometrical Journal*, 23(3):297–303.

[23] Tusell, F. (2007). estadística matemática. Technical report, Universidad del País Vasco.

[24] Velasco Vázquez, M. (2008). *Un Modelo de Regresió Poisson Inflado con Ceros para Analizar datos de un Experimento de Fungicidas en Jitomate*. PhD thesis, Universidad Veracruzana, Facultad de Estadística e Informática.

[25] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

[26] Winkelmann, R. (2008). *Econometric Analysis of Count Data*. Springer Berlin Heidelberg.

[27] Yee, T. W. (2015). *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer, New York, USA.

[28] Zeileis, A. (2004a). Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, 11(10):1–17.

[29] Zeileis, A. (2004b). Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, 11(1):1–17.

[30] Zeileis, A. and Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, 2(3):7–10.

[31] Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in r. *Journal of Statistical Software*, 27(1):1–25.

[32] Zha, L. (2014). The poisson inverse gaussian (pig) generalized linear regression model for analyzing motor vehicle crash data. *Zachry Department of Civil Engineering, Texas A&M University*.

80

Lista de Abreviaturas

(AIC) Akaike Information Criterion

(BIC) Bayesian Information Criterion

(BN) Binomial Negativo

(CPO) Cariado, Perdido, Obturado

(ECH) Encuesta Contínua de Hogares

(EMC) Estimador Mínimo Cuadrático

(EUA) Estados Unidos de América

(IESTA) Instituto de Estadística

(IG) Inversa Gaussiana

(INSE) Índice de Nivel Socio Económico

(MCI) Modelo Cero Inflado

(MCT) Modelo Cero Truncado

(MH) Modelo Hurdle

(MLG) Modelos Lineales Generalizados

(MRL) Modelos de Regresión Lineal

(MSP) Ministerio de Salud Pública

(MV) Máxima Verosimilitud

(N-R) Newton-Rapson

(OMS) Organización Mundial de la Salud

(PIG) Poisson Inversa Gaussiana

(Q-P) Quasi-Poisson

(SCE) Suma de Cuadrados Explicados

(SCR) Suma de Cuadrados de los Residuos

(SNIS) Sistema Nacional Integrado de Salud

(VA) Variable Aleatoria

82

Apéndice A

Script de variable Ccorona

```
load('datos_odonto.RData')

# CARGAMOS LIBRERIAS #

library(pscl)
library(sandwich)
library(lmtest)
library(MASS)
library(gamlss)
library(boot)
library(VGAM)
library(vcd)
library(xtable)
library(rcompanion)

#####################################################
## CREAMOS MATRIZ DE DATOS SIN DATOS FALTANTES ##

# SELECCIONAMOS LAS VARIABLES A USAR Y CONVERTIMOS EN NIVELES LAS QUE SON
    FACTORIALES #

datos=datos.odonto[,c(8,13,14,16,31,53,56,221,224,226,228,238,239)]
levels(datos$est_univers)=c("si","no",NA)
levels(datos$fuma)=c("si","no",NA)
datos$n5consumem=as.factor(datos$n5consumem)
levels(datos$n5consumem)=c("si","no",NA)

#################### CCORONA ####################

# SELECCIONAMOS LOS DATOS SIN FALTANTES #

a=complete.cases(datos[,c(1:8,13)])
regionCSF=datos$region[a]
tramo_etaCSF=datos$tramo_eta[a]
sexoCSF=datos$sexo[a]
est_universCSF=datos$est_univers[a]
institucinCSF=as.factor(datos$institucin[a])
n5consumemCSF=as.factor(datos$n5consumem[a])
fumaCSF=datos$fuma[a]
ccoronaCSF=datos$ccorona[a]
inseCSF=datos$inse[a]

# CREAMOS MATRIZ DE DATOS #
```

```
datosCSF=data.frame(ccoronaCSF,regionCSF,tramo_etaCSF,sexoCSF,est_universCSF,
institucinCSF,n5consumemCSF,fumaCSF,inseCSF)

# COMPROBAMOS QUE TODAS LAS VARIABLES TIENEN IGUAL DIMENSION #

summary(datosCSF)
length(datosCSF$regionCSF)
length(datosCSF$tramo_etaCSF)
length(datosCSF$sexoCSF)
length(datosCSF$est_universCSF)
length(datosCSF$institucinCSF)
length(datosCSF$n5consumemCSF)
length(datosCSF$fumaCSF)
length(datosCSF$ccoronaCSF)
length(datosCSF$inseCSF)

#####################################################
## VEO SI CPOCORONA = CCORONA+PCORONA+OCORONA ##

CPOc=datos$cpocorona[a]
Pc=datos$pcorona[a]
Oc=datos$ocorona[a]
sum(CPOc)
sum(ccoronaCSF)+sum(Pc)+sum(Oc)
#####################################################

###############################
## ESTADISTICA DESCRIPTIVA ##
###############################

# RESUMEN DE LAS VARIABLES #

summary(datosCSF)
summary(datosCSF$ccoronaCSF)
table(datosCSF$ccoronaCSF)

par(mfrow=c(1,2))
plot(table(datosCSF$ccoronaCSF),,ylab="Frecuencia", xlab="Ccorona",col="red")
boxplot(datosCSF$ccoronaCSF,xlab="Ccorona",ylab="")

# MEDIA Y VARIANZA #
mean(datosCSF$ccoronaCSF)
var(datosCSF$ccoronaCSF)

# HISTOGRAMA DE CCORONA CON SU DENSIDAD #
hist(datosCSF$ccoronaCSF,freq=F,ylim=c(0,0.7), ylab="f_Y(y)", xlab="Y",main="")
lines(density(datosCSF$ccoronaCSF),col='red')

#############################################################
## AJUSTE DE DISTRIBUCIONES A LA VARIABLE A EXPLICAR ##
#############################################################

## Ajustamos posibles distribuciones a la variable a explicar dada la naturaleza de
     los datos ##

# POISSON #
ycp=fitdistr(datosCSF$ccoronaCSF,"poisson")
lambda=ycp$estimate[1]), type="l", col="red", xlab="Y", ylab="f_Y(y)")
histDist(datosCSF$ccoronaCSF,"PO",density=TRUE, main="")

# BINOMIAL NEGATIVA #
ycbn=fitdistr(datosCSF$ccoronaCSF,"negative binomial")
histDist(datosCSF$ccoronaCSF,"NBII",density=TRUE, main="")
```

```
# PIG #
ycpig=gamlss(datosCSF$ccoronaCSF~1,family=PIG)
histDist(datosCSF$ccoronaCSF,"PIG",density=TRUE, main="")

# CERO INFL POISSON #
histDist(datosCSF$ccoronaCSF,"ZIP",density=TRUE,main="")

# CERO INFL BINOMIAL NEGATIVA #
histDist(datosCSF$ccoronaCSF,"ZINBI",density=TRUE, main="")

# CERO INFL PIG #
histDist(datosCSF$ccoronaCSF,"ZIPIG",density=TRUE, main="")

# HURDLE POISSON #
histDist(datosCSF$ccoronaCSF,"ZAP",density=TRUE, main="")

# HURDLE BINOMIAL NEGATIVA #
histDist(datosCSF$ccoronaCSF,"ZANBI",density=TRUE,main="")

# HURDLE PIG #
histDist(datosCSF$ccoronaCSF,"ZAPIG",density=TRUE, main="")

## ERRORES ##
## Calculamos errores absolutos y relativos de la distribucion

tablaCC=matrix(c(771,270,141,87,64,35,30,13,12,7,13,4,5,6,0,1,2,2,3), ncol=19, nrow
    =1, byrow=TRUE)

## Error absoluto
#suma de valor absoluto(y - y gorro)/n

#POIS#
nC=length(datosCSF$ccoronaCSF)

# Se calcula y gorro
probpoisC=dpois(min(datosCSF$ccoronaCSF):max(datosCSF$ccoronaCSF), lambda=
    ycp$estimate[1])
ygorpoisC=round(nC*probpoisC)

tablaCCP=c(tablaCC,ygorpoisC)

# BINOMIAL NEGATIVA
# y gorro
probbnC=dnbinom(min(datosCSF$ccoronaCSF):max(datosCSF$ccoronaCSF), size=
    ycbn$estimate[1], mu=ycbn$estimate[2])
ygorbnC=round(nC*probbnC)

tablaCCBN=c(tablaCC,ygorbnC)

### TEST DE BONDAD DE AJUSTE ###
# Pruebo que distribucion ajusta mejor #

## TEST CHI CUADRADO ##

# Poisson #
chiCpois=goodfit(datosCSF$ccoronaCSF, type="poisson", method="MinChisq")
summary(chiCpois)

# Binomial Negativo #
chiCbn=goodfit(datosCSF$ccoronaCSF, type="nbinomial", method="MinChisq")
summary(chiCbn)

##AIC##
```

```
AIC(ycp)
AIC(ycbn)

##BIC##
BIC(ycp)
BIC(ycbn)

#############################################################################

#####################################
#####     PROBAMOS MODELOS     ####
#####################################

# boxplot de ccorona y de ccorona con las variables explicativas #
par(mfrow=c(3,3))
boxplot(datosCSF$ccoronaCSF,xlab="Ccorona",border="dark red")
boxplot(datosCSF$ccoronaCSF~datosCSF$regionCSF, xlab="Region",border=c("dark blue
    ","dark green"))
boxplot(datosCSF$ccoronaCSF~datosCSF$tramo_etaCSF, xlab="Tramo etario",border=c("
    dark blue","dark green","purple"))
boxplot(datosCSF$ccoronaCSF~datosCSF$sexoCSF, xlab="Sexo",border=c("dark blue","
    dark green"))
boxplot(datosCSF$ccoronaCSF~datosCSF$est_universCSF,xlab="Estudio Universitario",
    border=c("dark blue","dark green"))
boxplot(datosCSF$ccoronaCSF~datosCSF$institucinCSF, xlab="Institucion Medica",
    border=c("dark blue","dark green"))
boxplot(datosCSF$ccoronaCSF~datosCSF$n5consumemCSF,xlab="Consume mate",border=c("
    dark blue","dark green"))
boxplot(datosCSF$ccoronaCSF~datosCSF$fumaCSF,xlab="Fuma",border=c("dark blue","dark
     green"))
plot(datosCSF$inseCSF,datosCSF$ccoronaCSF, xlab="INSE",cex=0.8,col="dark blue")

## Primer ajuste: BINOMIAL NEGATIVA ##

bnC=glm.nb(datosCSF$ccoronaCSF~.,data=datosCSF)
summary(bnC)

# Sacamos variables no significativas #
bnC2=glm.nb(datosCSF$ccoronaCSF~ datosCSF$est_universCSF+datosCSF$institucinCSF+
    datosCSF$n5consumemCSF+datosCSF$fumaCSF+
datosCSF$inseCSF,data=datosCSF)
summary(bnC2)

## Segundo ajuste: CERO INFLADO BINOMIAL NEGATIVA ##

ZIBNC=zeroinfl(datosCSF$ccoronaCSF~datosCSF$regionCSF+datosCSF$tramo_etaCSF+
    datosCSF$sexoCSF+datosCSF$est_universCSF+
datosCSF$institucinCSF+datosCSF$n5consumemCSF+datosCSF$fumaCSF+datosCSF$inseCSF|
    datosCSF$regionCSF+datosCSF$tramo_etaCSF+datosCSF$sexoCSF+
    datosCSF$est_universCSF+
datosCSF$institucinCSF+datosCSF$n5consumemCSF+datosCSF$fumaCSF+datosCSF$inseCSF,
    data=datosCSF,dist="negbin")
summary(ZIBNC)

## Tercer ajuste: HURDLE (CERO ALTERADO) BINOMIAL NEGATIVA ##

HURBNC=hurdle(datosCSF$ccoronaCSF~datosCSF$regionCSF+datosCSF$tramo_etaCSF+
    datosCSF$sexoCSF+datosCSF$est_universCSF+
datosCSF$institucinCSF+datosCSF$n5consumemCSF+datosCSF$fumaCSF+datosCSF$inseCSF|
    datosCSF$regionCSF+datosCSF$tramo_etaCSF+datosCSF$sexoCSF+
    datosCSF$est_universCSF+
datosCSF$institucinCSF+datosCSF$n5consumemCSF+datosCSF$fumaCSF+datosCSF$inseCSF,
    data=datosCSF,dist="negbin")
summary(HURBNC)
```

```
HURBNC2=hurdle(datosCSF$ccoronaCSF~datosCSF$est_universCSF+datosCSF$institucinCSF+
    datosCSF$fumaCSF+datosCSF$inseCSF|datosCSF$institucinCSF+datosCSF$n5consumemCSF
    +datosCSF$fumaCSF+datosCSF$inseCSF,data=datosCSF,dist="negbin")
summary(HURBNC2)

# Comparamos valores observados con valores esperados de Binomial Negativa y Hurdle
     Bin Neg

MATCH=matrix(c(0:18),ncol=3,nrow=19,dimnames=list(c(0:18),c("Observados","Binomial
    Negativa", "Hurdle")))
for (i in 1:19) {
A=round(c("Observados" = sum(datosCSF[,1] == i-1), "Binomial Negativa"=sum(dnbinom(
    i-1, mu = fitted(bnC2), size = bnC2$theta)),"Hurdle"=sum(predict(HURBNC2,type="
    prob")[,i])))
MATCH[i,1:3]=A
}
MATCH

################################
## VALIDACION Y DIAGNOSTICO ##
################################

par(mfrow=c(2,1))

# Validacion del modelo Binomial Negativo #

cov(bnC2$fitted,bnC2$residuals)
mean(bnC2$residuals)
plot(bnC2$residuals,ylim=c(-2,15), cex = 0.5,col="dark blue")
plot(residuals(bnC2)-fitted(bnC2),col="dark green")

# Validacion del modelo Hurdle Binomial Negativo #

cov(HURBNC2$fitted,bnC2$residuals)
mean(HURBNC2$residuals)
plot(HURBNC2$residuals,ylim=c(-2,15), cex = 0.5,col="dark blue")
plot(residuals(HURBNC2)-fitted(HURBNC2),col="dark green")

#####Error absoluto#####
#Suma de (valor absoluto de y - y estimado) sobre n

#Binomial Negativa#

ajusteBNC=fitted(bnC2)
ygormbnC=round(ajusteBNC)

errBNajusC=seq(1,1466)
for (i in 1:1466){
errBNajusC[i]=abs(datosCSF$ccoronaCSF[i]-ygormbnC[i])
}
ErrorBinomNegatC=sum(errBNajusC)/length(ygormbnC)

#pseudo-R2#
nagelkerke(bnC2)

AIC(bnC2)

#Hurdle BN#

ajusteHURC=fitted(HURBNC2)
ygormHC=round(ajusteHURC)

errHURajusC=seq(1,1466)
```

```
for (i in 1:1466){
errHURajusC[i]=abs(datosCSF$ccoronaCSF[i]-ygormHC[i])
}
ErrorHurdleC=sum(errHURajusC)/length(ygormHC)

#psudoR2#
ModC1=update(HURBNC2,.~1)
LIC=logLik(ModC1)
LFC=logLik(HURBNC2)
pR2C=1-(LFC/LIC)
pR2C

AIC(HURBNC2)

#Validacion cruzada
#BN#
#muestra
set.seed(71)
muestraC=sample(1:1466,1000,replace=FALSE)
muestrapruebaC=datosCSF[muestraC,]

#Binomial Negativa#

BNCvalid=glm.nb(muestrapruebaC$ccoronaCSF~muestrapruebaC$est_universCSF+
    muestrapruebaC$institucinCSF+muestrapruebaC$n5consumemCSF+
    muestrapruebaC$fumaCSF+
muestrapruebaC$inseCSF,data=muestrapruebaC)
summary(BNCvalid)
summary(bnC2)
#Cp=sum(round(predict(BNCvalid,newdata=muestrapruebaC,type="response"))==1)

#Hurdle Binomial Negativa#

HBNCvalid=hurdle(muestrapruebaC$ccoronaCSF~muestrapruebaC$est_universCSF+
    muestrapruebaC$institucinCSF+muestrapruebaC$fumaCSF+muestrapruebaC$inseCSF|
    muestrapruebaC$institucinCSF+muestrapruebaC$n5consumemCSF+
    muestrapruebaC$fumaCSF+muestrapruebaC$inseCSF,dist="negbin",data=muestrapruebaC
    )
summary(HBNCvalid)
summary(HURBNC2)

MATCval=matrix(c(0:18),ncol=3,nrow=19,dimnames=list(c(0:18),c("Observados","
    Binomial Negativa Validacion","Hurdle Validacion")))
for (i in 1:19) {
Aval=round(c("Observados"=sum(datosCSF[muestraC,1]==i-1),"Binomial Negativa V"=
    sum(dnbinom(i-1, mu = fitted(BNCvalid), size = BNCvalid$theta)),"Hurdle V"=sum(
    predict(HBNCvalid,type="prob")[,i])))
MATCval[i,1:3]=Aval
}
MATCval

#Prediccion

muestrapruebaC=datosCSF[-c(muestraC),]

MATCpred=matrix(c(0:18),ncol=3,nrow=19,dimnames=list(c(0:18),c("Observados","
    Binomial Negativa Prediccion","Hurdle Prediccion")))
for (i in 1:19) {
Apred=round(c("Observados"=sum(datosCSF[-c(muestraC),1]==i-1),"Binom Neg P"=sum
    (round(predict(BNCvalid,newdata=muestrapruebaC,type="response"))==i-1),"HURDLE
    P"=sum(predict(HBNCvalid, newdata=muestrapruebaC,type="prob")[,i])))
MATCpred[i,1:3]=Apred
}
MATCpred
```

Apéndice B

Script de variable Pcorona

```
load('datos_odonto.RData')

# CARGAMOS LIBRERIAS #

library(pscl)
library(sandwich)
library(lmtest)
library(MASS)
library(gamlss)
library(boot)
library(VGAM)
library(vcd)
library(xtable)
library(rcompanion)

#####################################################
## CREAMOS MATRIZ DE DATOS SIN DATOS FALTANTES ##

# SELECCIONAMOS LAS VARIABLES A USAR Y CONVERTIMOS EN NIVELES LAS QUE SON
    FACTORIALES #

datos=datos.odonto[,c(8,13,14,16,31,53,56,221,224,226,228,238,239)]
levels(datos$est_univers)=c("si","no",NA)
levels(datos$fuma)=c("si","no",NA)
datos$n5consumem=as.factor(datos$n5consumem)
levels(datos$n5consumem)=c("si","no",NA)

#################### PCORONA ####################

# SELECCIONAMOS LOS DATOS SIN FALTANTES #

b=complete.cases(datos[,c(1:7,9,13)])
regionPSF=datos$region[b]
tramo_etaPSF=datos$tramo_eta[b]
sexoPSF=datos$sexo[b]
est_universPSF=datos$est_univers[b]
institucinPSF=as.factor(datos$institucin[b])
n5consumemPSF=as.factor(datos$n5consumem[b])
fumaPSF=datos$fuma[b]
pcoronaPSF=datos$pcorona[b]
insePSF=datos$inse[b]

# CREAMOS MATRIZ DE DATOS #
```

```
datosPSF=data.frame(pcoronaPSF,regionPSF,tramo_etaPSF,sexoPSF,est_universPSF,
institucinPSF,n5consumemPSF,fumaPSF,insePSF)

#El 32 es un problema, se eliminan de la matriz de datos#
datosPSF=datosPSF[-which(32==datosPSF$pcorona),]

# COMPROBAMOS QUE TODAS LAS VARIABLES TIENEN IGUAL DIMENSION #

summary(datosPSF)
length(datosPSF$regionPSF)
length(datosPSF$tramo_etaPSF)
length(datosPSF$sexoPSF)
length(datosPSF$est_universPSF)
length(datosPSF$institucinPSF)
length(datosPSF$n5consumemPSF)
length(datosPSF$fumaPSF)
length(datosPSF$pcoronaPSF)
length(datosPSF$insePSF)

################################
## ESTADISTICA DESCRIPTIVA ##
################################

# RESUMEN DE LAS VARIABLES #
par(mfrow=c(1,2))
summary(datosPSF)
summary(datosPSF$pcoronaPSF)
table(datosPSF$pcoronaPSF)
plot(table(datosPSF$pcoronaPSF),col="dark red",ylab="Frecuencia", xlab="Pcorona")
boxplot(datosPSF$pcoronaPSF,xlab="Pcorona",ylab=" ")

# MEDIA Y VARIANZA #
mean(datosPSF$pcoronaPSF)
var(datosPSF$pcoronaPSF)

# HISTOGRAMA DE CCORONA CON SU DENSIDAD #
hist(datosPSF$pcoronaPSF,freq=F,ylim=c(0,0.3))
lines(density(datosPSF$pcoronaPSF),col='red')

############################################################
## AJUSTE DE DISTRIBUCIONES A LA VARIABLE A EXPLICAR ##
############################################################

## Ajustamos posibles distribuciones a la variable a explicar dada la naturaleza de
     los datos ##
par(mfrow=c(1,3))

#POISSON#
ypp=fitdistr(datosPSF$pcoronaPSF,"poisson")
histDist(datosPSF$pcoronaPSF,"PO",density=TRUE, main="Ajuste Poisson")

# BINOMIAL NEGATIVA #
ypbn=fitdistr(datosPSF$pcoronaPSF,"negative binomial")
histDist(datosPSF$pcoronaPSF,"NBII",density=TRUE,main="Ajuste Binomial Negativo")

# PIG #
yppig=gamlss(datosPSF$pcoronaPSF~1,family=PIG)
histDist(datosPSF$pcoronaPSF,"PIG",density=TRUE,main="Ajuste PIG")

#CERO INFL POISSON#
histDist(datosPSF$pcoronaPSF,"ZIP",density=TRUE)

#CERO INFL BINOMIAL NEGATIVA#
```

```
histDist(datosPSF$pcoronaPSF,"ZINBI",density=TRUE,main="")

#CERO INFL PIG#
histDist(datosPSF$pcoronaPSF,"ZIPIG",density=TRUE)

#HURDLE POISSON#
histDist(datosPSF$pcoronaPSF,"ZAP",density=TRUE)

#HURDLE BINOMIAL NEGATIVA#
histDist(datosPSF$pcoronaPSF,"ZANBI",density=TRUE,main="")

# HURDLE PIG #
histDist(datosPSF$pcoronaPSF,"ZAPIG",density=TRUE, main="")

##############################################################################

#####################################
#####    PROBAMOS MODELOS    ####
#####################################

# boxplot de ccorona y de ccorona con las variables explicativas #
par(mfrow=c(3,3))
boxplot(datosPSF$pcoronaPSF,xlab="Pcorona", border="dark red")
boxplot(datosPSF$pcoronaPSF~datosPSF$regionPSF, xlab="Region", border=c("dark blue
    ","dark green"))
boxplot(datosPSF$pcoronaPSF~datosPSF$tramo_etaPSF, xlab="Tramo etario", border=c("
    dark blue","dark green","purple"))
boxplot(datosPSF$pcoronaPSF~datosPSF$sexoPSF, xlab="Sexo",border=c("dark blue","
    dark green"))
boxplot(datosPSF$pcoronaPSF~datosPSF$est_universPSF,xlab="Estudio Universitario",
    border=c("dark blue","dark green"))
boxplot(datosPSF$pcoronaPSF~datosPSF$institucinPSF, xlab="Institucion Medica",
    border=c("dark blue","dark green"))
boxplot(datosPSF$pcoronaPSF~datosPSF$n5consumemPSF,xlab="Consume mate",border=c("
    dark blue","dark green"))
boxplot(datosPSF$pcoronaPSF~datosPSF$fumaPSF,xlab="Fuma",border=c("dark blue","dark
     green"))
plot(datosPSF$insePSF,datosPSF$pcoronaPSF, xlab="INSE",ylab="",cex=0.7,col="dark
    blue")

## Primer ajuste: CERO INFLADO BINOMIAL NEGATIVA ##

ZIBNP=zeroinfl(datosPSF$pcoronaPSF~datosPSF$regionPSF+datosPSF$tramo_etaPSF+
    datosPSF$sexoPSF+datosPSF$est_universPSF+
datosPSF$institucinPSF+datosPSF$n5consumemPSF+datosPSF$fumaPSF+datosPSF$insePSF|
    datosPSF$regionPSF+datosPSF$tramo_etaPSF+datosPSF$sexoPSF+
    datosPSF$est_universPSF+
datosPSF$institucinPSF+datosPSF$n5consumemPSF+datosPSF$fumaPSF+datosPSF$insePSF,
    data=datosPSF,dist="negbin")
summary(ZIBNP)

# Sacamos variables no significativas #
ZIBNP2=zeroinfl(datosPSF$pcoronaPSF~datosPSF$tramo_etaPSF+datosPSF$sexoPSF+
datosPSF$institucinPSF+datosPSF$fumaPSF+datosPSF$insePSF|datosPSF$regionPSF+
    datosPSF$tramo_etaPSF+datosPSF$est_universPSF+
datosPSF$n5consumemPSF+datosPSF$fumaPSF+datosPSF$insePSF,data=datosPSF,dist="negbin
    ")
summary(ZIBNP2)

## Segundo ajuste: HURDLE (CERO ALTERADO) BINOMIAL NEGATIVA ##

HURBNP=hurdle(datosPSF$pcoronaPSF~datosPSF$regionPSF+datosPSF$tramo_etaPSF+
    datosPSF$sexoPSF+datosPSF$est_universPSF+
datosPSF$institucinPSF+datosPSF$n5consumemPSF+datosPSF$fumaPSF+datosPSF$insePSF|
```

```
    datosPSF$regionPSF+datosPSF$tramo_etaPSF+datosPSF$sexoPSF+
    datosPSF$est_universPSF+
datosPSF$institucinPSF+datosPSF$n5consumemPSF+datosPSF$fumaPSF+datosPSF$insePSF,
    data=datosPSF,dist="negbin")
summary(HURBNP)

#Sacamos variable no significativas#

HURBNP2=hurdle(datosPSF$pcoronaPSF~datosPSF$tramo_etaPSF+datosPSF$sexoPSF+
datosPSF$institucinPSF+datosPSF$fumaPSF+datosPSF$insePSF|datosPSF$regionPSF+
    datosPSF$tramo_etaPSF+datosPSF$est_universPSF+
datosPSF$n5consumemPSF+datosPSF$fumaPSF+datosPSF$insePSF,data=datosPSF,dist="negbin
    ")
summary(HURBNP2)

# Comparamos valores observados con valores esperados de Cero Inflado Binomial
    Negativa y Hurdle Binomial Negativa

MATPE=matrix(c(0:31),ncol=3,nrow=32,dimnames=list(c(0:31),c("Observados","Zero
    Inflado","Hurdle")))
for (i in 1:32) {
A=round(c("Observados" = sum(datosPSF[,1] == i-1), "ZIBN"=sum(predict(ZIBNP2,type="
    prob")[,i]),"HURDLE"=sum(predict(HURBNP2,type="prob")[,i])))
MATPE[i,1:3]=A
}
MATPE

################################
## VALIDACION Y DIAGNOSTICO ##
################################

par(mfrow=c(2,1))

# Validacion del modelo Cero Inflado Binomial Negativo #

cov(ZIBNP2$fitted,ZIBNP2$residuals)
mean(ZIBNP2$residuals)
plot(ZIBNP2$residuals, cex = 0.5,col="dark blue",ylab="Residuos",xlab="")
plot(residuals(ZIBNP2)-fitted(ZIBNP2),col="dark green",ylab="Residuos vs. Ajustados
    ",xlab="")

# Validacion del modelo Hurdle Binomial Negativo #

cov(HURBNP2$fitted,HURBNP2$residuals)
mean(HURBNP2$residuals)
plot(HURBNP2$residuals, cex = 0.5,col="dark blue",ylab="Residuos",xlab="")
plot(residuals(HURBNP2)-fitted(HURBNP2), col="dark green",ylab="Residuos vs.
    Ajustados",xlab="")

#####Error absoluto#####
#Suma de (valor absoluto de y - y estimado) sobre n

#Cero Inflado#

ygorZIBNP2=round(predict(ZIBNP2))
nP=length(datosPSF$pcoronaPSF)

errZIBN2=seq(1:1350)
for (i in 1:nP) {
errZIBN2[i]=abs(datosPSF$pcoronaPSF[i]-ygorZIBNP2[i])
}
ErrorZIBinNegP=sum(errZIBN2)/nP

#psudoR2#
```

```
ModP1ZI=update(ZIBNP2,.~1)
LIPZI=logLik(ModP1ZI)
LFPZI=logLik(ZIBNP2)
pR2PZI=1-(LFPZI/LIPZI)
pR2PZI

AIC(ZIBNP2)

#Hurdle#

ygorHURBNP2=round(predict(HURBNP2))
nP=length(datosPSF$pcoronaPSF)

errHURBN2=seq(1:1350)
for (i in 1:nP) {
errHURBN2[i]=abs(datosPSF$pcoronaPSF[i]-ygorHURBNP2[i])
}
ErrorHBinNegP=sum(errHURBN2)/nP

#psudoR2#
ModP1H=update(HURBNP2,.~1)
LIPH=logLik(ModP1H)
LFPH=logLik(HURBNP2)
pR2PH=1-(LFPH/LIPH)
pR2PH

AIC(HURBNP2)

##Validacion cruzada##

#muestra
set.seed(21)
muestraP=sample(1:1350,1000,replace=FALSE)
muestrapruebaP=datosPSF[muestraP,]

#Cero Inflado Binomial Negativa#

ZIBNPvalid=zeroinfl(muestrapruebaP$pcoronaPSF~muestrapruebaP$tramo_etaPSF+
    muestrapruebaP$sexoPSF+
muestrapruebaP$institucinPSF+muestrapruebaP$fumaPSF+muestrapruebaP$insePSF|
    muestrapruebaP$regionPSF+muestrapruebaP$tramo_etaPSF+
muestrapruebaP$est_universPSF+muestrapruebaP$n5consumemPSF+muestrapruebaP$fumaPSF+
    muestrapruebaP$insePSF,data=muestrapruebaP,dist="negbin")
summary(ZIBNPvalid)
summary(ZIBNP2)

#Hurdle Binomial Negativa#

HURBNPvalid=hurdle(muestrapruebaP$pcoronaPSF~muestrapruebaP$tramo_etaPSF+
    muestrapruebaP$sexoPSF+muestrapruebaP$institucinPSF+muestrapruebaP$fumaPSF+
muestrapruebaP$insePSF|muestrapruebaP$regionPSF+muestrapruebaP$tramo_etaPSF+
    muestrapruebaP$est_universPSF+muestrapruebaP$n5consumemPSF+
    muestrapruebaP$fumaPSF+muestrapruebaP$insePSF,data=muestrapruebaP,dist="negbin
    ")
summary(HURBNPvalid)
summary(HURBNP2)

MATPval=matrix(c(0:31),ncol=3,nrow=32,dimnames=list(c(0:31),c("Observados", "Hurdle
     Validacion","Cero Inflado Validacion")))
for (i in 1:32) {
Aval=round(c("Observados" = sum(datosPSF[muestraP,1] == i-1),"HURDLE V"=sum(predict
    (HURBNPvalid,type="prob")[,i]),"CERO INFLADO V"=sum(predict(ZIBNPvalid,type="
    prob")[,i])))
MATPval[i,1:3]=Aval
```

```
}
MATPval

#Prediccion

muestrapruebaP=datosPSF[-c(muestraP),]

MATPpred=matrix(c(0:31),ncol=3,nrow=32,dimnames=list(c(0:31),c("Observados", "Cero
    Inflado Prediccion","Hurdle Prediccion")))
for (i in 1:32) {
Apred=round(c("Observados" = sum(datosPSF[-c(muestraP),1] == i-1),"CERO INFL P"=sum
    (predict(ZIBNPvalid, newdata=muestrapruebaP,type="prob")[,i]),"HURDLE P"=sum(
    predict(HURBNPvalid, newdata=muestrapruebaP,type="prob")[,i])))
MATPpred[i,1:3]=Apred
}
MATPpred
```

Apéndice C

Script de variable Ocorona

```
load('datos_odonto.RData')

library(pscl)
library(sandwich)
library(lmtest)
library(MASS)
library(gamlss)
library(xtable) #para las tablas en chrome
library(rcompanion) #para el pseudo R2

##################################################
## CREAMOS MATRIZ DE DATOS SIN DATOS FALTANTES ##

# SELECCIONAMOS LAS VARIABLES A USAR Y CONVERTIMOS EN NIVELES LAS QUE SON
    FACTORIALES #

datos=datos.odonto[,c(8,13,14,16,31,53,56,221,224,226,228,238,239)]
levels(datos$est_univers)=c("si","no",NA)
levels(datos$fuma)=c("si","no",NA)
datos$n5consumem=as.factor(datos$n5consumem)
levels(datos$n5consumem)=c("si","no",NA)

################## OCORONA ####################

# SELECCIONAMOS LOS DATOS SIN FALTANTES #

c=complete.cases(datos[,c(1:7,10,12)])
regionOSF=datos$region[c]
tramo_etaOSF=datos$tramo_eta[c]
sexoOSF=datos$sexo[c]
est_universOSF=datos$est_univers[c]
institucinOSF=as.factor(datos$institucin[c])
n5consumemOSF=as.factor(datos$n5consumem[c])
fumaOSF=datos$fuma[c]
ocoronaOSF=as.numeric(datos$ocorona[c])
inseOSF=as.numeric(datos$inse[c])

# CREAMOS MATRIZ DEDATOS #

datosOSF=data.frame(ocoronaOSF,regionOSF,tramo_etaOSF,sexoOSF,est_universOSF,
institucinOSF,n5consumemOSF,fumaOSF,inseOSF)

# COMPROBAMOS QUE TODAS LAS VARIABLES TIENEN IGUAL DIMENSION #
```

```
summary(datosOSF)
length(datosOSF$regionOSF)
length(datosOSF$tramo_etaOSF)
length(datosOSF$sexoOSF)
length(datosOSF$est_universOSF)
length(datosOSF$institucinOSF)
length(datosOSF$n5consumemOSF)
length(datosOSF$fumaOSF)
length(datosOSF$ocoronaOSF)
length(datosOSF$inseOSF)

################################
## ESTADISTICA DESCRIPTIVA ##
################################

# RESUMEN DE LAS VARIABLES #
par(mfrow=c(1,2))
summary(datosOSF)
summary(datosOSF$ocoronaOSF)
table(datosOSF$ocoronaOSF)
plot(table(datosOSF$ocoronaOSF),col="red",xlab="Ocorona",ylab="Frecuencia")
boxplot(datosOSF$ocoronaOSF,xlab="Ocorona",ylab="")

# MEDIA Y VARIANZA #
mean(datosOSF$ocoronaOSF)

var(datosOSF$ocoronaOSF)

# HISTOGRAMA DE OCORONA CON SU DENSIDAD #
hist(datosOSF$ocoronaOSF,freq=F,ylim=c(0,0.4), ylab="f_Y(y)", xlab="Y",main="")
lines(density(datosOSF$ocoronaOSF),col='red')

############################################################
## AJUSTE DE DISTRIBUCIONES A LA VARIABLE A EXPLICAR ##
############################################################

## Ajustamos posibles distribuciones a la variable a explicar dada la naturaleza de
     los datos ##

# POISSON #
yop=fitdistr(datosOSF$ocoronaOSF,"poisson")
histDist(datosOSF$ocoronaOSF,"PO",density=TRUE,,ylab="",main="Ajuste Poisson")

# BINOMIAL NEGATIVA #
yobn=fitdistr(datosOSF$pcoronaOSF,"negative binomial")
histDist(datosOSF$ocoronaOSF,"NBII",density=TRUE,,ylab="",main="Ajuste Binomial
    Negativo")

# PIG #
histDist(datosOSF$ocoronaOSF,"PIG",density=TRUE,,ylab="",main="Ajuste PIG")

# CERO INFL POISSON #
histDist(datosOSF$ocoronaOSF,"ZIP",density=TRUE)

# CERO INFL BINOMIAL NEGATIVA #
histDist(datosOSF$ocoronaOSF,"ZINBI",density=TRUE,main="")

# CERO INFLADO PIG #
histDist(datosOSF$ocoronaOSF,"ZIPIG",density=TRUE)

# HURDLE POISSON #
histDist(datosOSF$ocoronaOSF,"ZAP",density=TRUE)
```

```
# HURDLE BINOMIAL NEGATIVA #
histDist(datosOSF$ocoronaOSF,"ZANBI",density=TRUE,main="")

#############################################################################

#####################################
#####     PROBAMOS MODELOS      ####
#####################################

# boxplot de ocorona y de ocorona con las variables explicativas #
par(mfrow=c(3,3))
boxplot(datosOSF$ocoronaOSF,border="dark red",xlab="Ocorona")
boxplot(datosOSF$ocoronaOSF~datosOSF$regionOSF,border=c("dark blue","dark green"),
    xlab="Region")
boxplot(datosOSF$ocoronaOSF~datosOSF$tramo_etaOSF,border=c("dark blue","dark green
    ","purple"),xlab="Tramo Etario")
boxplot(datosOSF$ocoronaOSF~datosOSF$sexoOSF,border=c("dark blue","dark green"),
    xlab="Sexo")
boxplot(datosOSF$ocoronaOSF~datosOSF$est_universOSF,border=c("dark blue","dark
    green"),xlab="Estudio Universitario")
boxplot(datosOSF$ocoronaOSF~datosOSF$institucinOSF,border=c("dark blue","dark green
    "),xlab="Institucion Medica")
boxplot(datosOSF$ocoronaOSF~datosOSF$n5consumemOSF,border=c("dark blue","dark green
    "),xlab="Consume Mate")
boxplot(datosOSF$ocoronaOSF~datosOSF$fumaOSF,border=c("dark blue","dark green"),
    xlab="Fuma")
plot(datosOSF$ocoronaOSF~datosOSF$inseOSF,cex = 0.6,col="dark blue",xlab="INSE")

## Primer ajuste: BINOMIAL NEGATIVA CERO INFLADO ##

ZIBNO=zeroinfl(datosOSF$ocoronaOSF~datosOSF$regionOSF+datosOSF$tramo_etaOSF+
    datosOSF$sexoOSF+datosOSF$est_universOSF+
datosOSF$institucinOSF+datosOSF$n5consumemOSF+datosOSF$fumaOSF+datosOSF$inseOSF|
    datosOSF$regionOSF+datosOSF$tramo_etaOSF+datosOSF$sexoOSF+
    datosOSF$est_universOSF+
datosOSF$institucinOSF+datosOSF$n5consumemOSF+datosOSF$fumaOSF+datosOSF$inseOSF,
    data=datosOSF,dist="negbin")
summary(ZIBNO)

# Sacamos variables no significativas #

ZIBNO2=zeroinfl(datosOSF$ocoronaOSF~datosOSF$tramo_etaOSF+datosOSF$sexoOSF+
    datosOSF$institucinOSF+
datosOSF$inseOSF|datosOSF$tramo_etaOSF+datosOSF$sexoOSF+datosOSF$est_universOSF+
    datosOSF$institucinOSF+
datosOSF$fumaOSF+datosOSF$inseOSF,data=datosOSF,dist="negbin")
summary(ZIBNO2)

## Segundo ajuste: BINOMIAL NEGATIVA HURDLE (CERO ALTERADO) ##

HURBNO=hurdle(datosOSF$ocoronaOSF~datosOSF$regionOSF+datosOSF$tramo_etaOSF+
datosOSF$sexoOSF+datosOSF$est_universOSF+datosOSF$institucinOSF+
datosOSF$n5consumemOSF+datosOSF$fumaOSF+datosOSF$inseOSF|datosOSF$regionOSF+
datosOSF$tramo_etaOSF+datosOSF$sexoOSF+datosOSF$est_universOSF+
    datosOSF$institucinOSF+
datosOSF$n5consumemOSF+datosOSF$fumaOSF+datosOSF$inseOSF,data=datosOSF,dist="negbin
    ")
summary(HURBNO)

HURBNO2=hurdle(datosOSF$ocoronaOSF~datosOSF$tramo_etaOSF+datosOSF$sexoOSF+
    datosOSF$institucinOSF+
datosOSF$inseOSF|datosOSF$tramo_etaOSF+datosOSF$sexoOSF+datosOSF$est_universOSF+
datosOSF$institucinOSF+datosOSF$fumaOSF+datosOSF$inseOSF,data=datosOSF,dist="negbin
    ")
```

```
summary(HURBNO2)

# Comparamos valores observados con valores esperados de Cero Inflado Binomial
    Negativa y Hurdle Binomial Negativa

MATO=matrix(c(0:31),ncol=3,nrow=32,dimnames=list(c(0:31),c("Observados","Zero
    Inflado", "Hurdle")))
for (i in 1:32) {
A=round(c("Observados" = sum(datosOSF[,1] == i-1), "ZIBN"=sum(predict(ZIBNO2,type="
    prob")[,i]),"HURDLE"=sum(predict(HURBNO2,type="prob")[,i])))
MATO[i,1:3]=A
}
MATO

##################################
## VALIDACION Y DIAGNOSTICO ##
##################################

par(mfrow=c(2,1))

# Validacion del modelo Cero Inflado Binomial Negativo #

cov(ZIBNO2$fitted,ZIBNO2$residuals)
mean(ZIBNO2$residuals)
plot(ZIBNO2$residuals, cex = 0.5,col="dark blue",xlab="",ylab="Residuos")
plot(residuals(ZIBNO2)-fitted(ZIBNO2),col="dark green",xlab="",ylab="Residuos vs
    Ajustados")

par(mfrow=c(2,1))

# Validacion del modelo Hurdle Binomial Negativo #

cov(HURBNO2$fitted,HURBNO2$residuals)
mean(HURBNO2$residuals)
plot(HURBNO2$residuals, cex = 0.5,col="dark blue",xlab="",ylab="Residuos")
plot(residuals(HURBNO2)-fitted(HURBNO2),col="dark green",xlab="",ylab="Residuos vs
    Ajustados")

# Errores de prediccion #
# Error absoluto
# Suma de (valor absoluto de y - y estimado) sobre n

# Cero Inflado #
ygorZIBNO2=round(predict(ZIBNO2))
nO=length(datosOSF$ocoronaOSF)

errZIBNO2=seq(0,1468)
for (i in 1:nO){
errZIBNO2[i]=abs(datosOSF$ocoronaOSF[i]-ygorZIBNO2[i])
}
sum(errZIBNO2)/nO

#psudoR2#
ModO1ZI=update(ZIBNO2,.~1)
LIOZI=logLik(ModO1ZI)
LFOZI=logLik(ZIBNO2)
pR2OZI=1-(LFOZI/LIOZI)
pR2OZI

AIC(LFOZI)

# Hurdle #

ygorHURBNO2=round(predict(HURBNO2))
```

```
nO=length(datosOSF$ocoronaOSF)

errHURBNO2=seq(0,1468)
for (i in 1:1469){
errHURBNO2[i]=abs(datosOSF$ocoronaOSF[i]-ygorHURBNO2[i])
}
sum(errHURBNO2)/nO

# Pseudo R2 para Cero Inflado
nagelkerke(ZIBNO2)

#psudoR2#
ModO1H=update(HURBNO2,.~1)
LIOH=logLik(ModO1H)
LFOH=logLik(HURBNO2)
pR2OH=1-(LFOH/LIOH)
pR2OH

AIC(HURBNO2)

## Validacion cruzada ##

#muestra
set.seed(512)
muestraO=sample(1:1350,1000,replace=FALSE)
muestrapruebaO=datosOSF[muestraO,]

# Hurdle #
BNHUROvalid=hurdle(muestrapruebaO$ocoronaOSF~muestrapruebaO$tramo_etaOSF+
    muestrapruebaO$sexoOSF+muestrapruebaO$institucinOSF+
muestrapruebaO$inseOSF|muestrapruebaO$tramo_etaOSF+muestrapruebaO$sexoOSF+
    muestrapruebaO$est_universOSF+
muestrapruebaO$institucinOSF+muestrapruebaO$fumaOSF+muestrapruebaO$inseOSF,data=
    muestrapruebaO,dist="negbin")
summary(BNHUROvalid)

# Cero Inflado #
BNZIOvalid=zeroinfl(muestrapruebaO$ocoronaOSF~muestrapruebaO$tramo_etaOSF+
    muestrapruebaO$sexoOSF+muestrapruebaO$institucinOSF+
muestrapruebaO$inseOSF|muestrapruebaO$tramo_etaOSF+muestrapruebaO$sexoOSF+
    muestrapruebaO$est_universOSF+
muestrapruebaO$institucinOSF+muestrapruebaO$fumaOSF+muestrapruebaO$inseOSF,data=
    muestrapruebaO,dist="negbin")
summary(BNZIOvalid)

# Observados vs. Predichos #

MATOval=matrix(c(0:31),ncol=3,nrow=32,dimnames=list(c(0:31),c("Observados", "Hurdle
    Validacion", "Cero Inflado Validacion")))
for (i in 1:32) {
Aval=round(c("Observados" = sum(datosOSF[muestraO,1] == i-1),"HURDLE V"=sum(predict
    (BNHUROvalid,type="prob")[,i]),"ZERO INF V"=sum(predict(BNZIOvalid,type="prob")
    [,i])))
MATOval[i,1:3]=Aval
}
MATOval

# Prediccion #

muestrapruebaO=datosOSF[-c(muestraO),]

MATOpred=matrix(c(0:31),ncol=3,nrow=32,dimnames=list(c(0:31),c("Observados", "Cero
    Inflado Prediccion","Hurdle Prediccion")))
for (i in 1:32) {
```

```
Apred=round(c("Observados" = sum(datosOSF[-c(muestraO),1] == i-1),"CERO INF P"=sum(
    predict(BNZIOvalid, newdata=muestrapruebaO,type="prob")[,i]),"Hurdle P"=sum(
    predict(BNHUROvalid, newdata=muestrapruebaO,type="prob")[,i])))
MATOpred[i,1:3]=Apred
}
MATOpred
```

Apéndice D

Script de variable CPOcorona

```
load('datos_odonto.RData')

library(pscl)
library(sandwich)
library(lmtest)
library(MASS)
library(gamlss)

##################################################
## CREAMOS MATRIZ DE DATOS SIN DATOS FALTANTES ##

datos=datos.odonto[,c(8,13,14,16,31,53,56,221,224,226,228,238,239)]
levels(datos$est_univers)=c("si","no",NA)
levels(datos$fuma)=c("si","no",NA)

################## CPOCORONA ####################

## Seleccionamos los datos sin faltantes
a=complete.cases(datos[,c(1:7,11,12)])
regionSF=datos$region[a]
tramo_etaSF=datos$tramo_eta[a]
sexoSF=datos$sexo[a]
est_universSF=datos$est_univers[a]
institucinSF=as.factor(datos$institucin[a])
n5consumemSF=as.factor(datos$n5consumem[a])
fumaSF=datos$fuma[a]
cpocoronaSF=as.numeric(datos$cpocorona[a])
inseSF=as.numeric(datos$inse[a])

## Creamos matriz de datos
datosSF=data.frame(cpocoronaSF,regionSF,tramo_etaSF,sexoSF,est_universSF,
institucinSF,n5consumemSF,fumaSF,inseSF)

#El 32 es un problema#
datosSF=datosSF[-which(32==datosSF$cpocorona),]

## Comprobamos que todas las variables tienen igual dimension
summary(datosSF)
length(datosSF$regionSF)
length(datosSF$tramo_etaSF)
length(datosSF$sexoSF)
length(datosSF$est_universSF)
length(datosSF$institucinSF)
```

```
length(datosSF$n5consumemSF)
length(datosSF$fumaSF)
length(datosSF$cpocoronaSF)
length(datosSF$inseSF)

################################
## ESTADISTICA DESCRIPTIVA ##
################################

summary(datosSF)
summary(datosSF$cpocoronaSF)
table(datosSF$cpocoronaSF)
plot(table(datosSF$cpocoronaSF))

mean(datosSF$cpocoronaSF)
var(datosSF$cpocoronaSF)

hist(datosSF$cpocoronaSF,freq=F,breaks=31,border="dark blue",col="light green",main
    ="",xlab="",ylab="")
lines(density(datosSF$cpocoronaSF),col='red')

###################################
##DISTRIBUCIONES: POI, BN, PIG##
###################################

## Ajustamos distribuciones a la variable a explicar

#POIS#
ycpop=fitdistr(datosSF$cpocoronaSF,"poisson")
histDist(datosSF$cpocoronaSF,"PO",density=TRUE)

#BN#
ycpobn=fitdistr(datosSF$cpocoronaSF,"negative binomial")
histDist(datosSF$cpocoronaSF,"NBII",density=TRUE)

#PIG##
ycpopig=gamlss(datosSF$cpocoronaSF~1,family=PIG)
histDist(datosSF$cpocoronaSF,"PIG",density=TRUE)

#CERO INFL POISSON#
histDist(datosSF$cpocoronaSF,"ZIP",density=TRUE)

#CERO INFL NB#
histDist(datosSF$cpocoronaSF,"ZINBI",density=TRUE)

#CERO INFL PIG#
histDist(datosSF$cpocoronaSF,"ZIPIG",density=TRUE)

#HURDLE POISSON#
histDist(datosSF$cpocoronaSF,"ZAP",density=TRUE)

#HURDLE NB#
histDist(datosSF$cpocoronaSF,"ZANBI",density=TRUE)

#HURDLE PIG#
histDist(datosSF$cpocoronaSF,"ZAPIG",density=TRUE)
```

Apéndice E

Diseño y selección de la muestra

El diseño de la muestra fue realizado por el Servicio de Epidemiología de la Cátedra de Odontología Social en colaboración con el Instituto de Estadística (IESTA) de la Facultad de Ciencias Económicas y de Administración. Los tramos etarios fueron seleccionados de acuerdo al siguiente criterio:

- 15 a 24 años: es la edad que la OMS recomienda para relevar la situación epidemiológica en la salud bucal de los jóvenes.
- 35 a 44 años: permite conocer tanto el estado de salud bucal de los adultos como los efectos de los tratamientos que han recibido hasta el momento, además de que es la edad recomendada por la OMS para realizar comparaciones internacionales.
- 65 a 74 años: permite conocer los efectos de los tratamientos recibidos por los adultos mayores, además de la importancia que cobra al ser la población uruguaya una de las que presenta mayor proporción de población adulta en Latinoamérica ($4,5\,\%$)

Se trabajó con 2 muestras independientes: por un lado se consideraron los departamentos del interior del país y por el otro Montevideo.

El diseño muestral se realizó en 2 fases:

- En la primera fase el marco muestral fue el conjunto de personas de los 3 tramos etarios pertenecientes a localidades de 20.000 o más habitantes que fueron visitadas en la Encuesta Continua de Hogares (ECH) desde febrero a abril de 2010 [1] (ECH *http://ine.gub.uy/encuesta-continua-de-hogares1*)

[1] Instituto Nacional de Estadística, División Estadísticas Sociodemográficas, Departamento Encuesta de Hogares; Inicio: 1968 - Actualmente en ejecución.

- En la segunda etapa se realiza una muestra del total de personas de la primera etapa y se llega así al total requerido.

El tamaño muestral fue calculado de la siguiente forma:

$$n = \left[\frac{(\phi_{1-\alpha/2})^2 * \pi * (1-\pi)}{(Moe)^2}\right] * Deff * \left\{\frac{1}{1-TNR}\right\} \tag{E.1}$$

Donde Moe es el margen de error deseado, $Deff$ es el efecto diseño (inflación de varianza por muestreo complejo), TNR es la tasa de no respuesta, π es la prevalencia a ser estimada y ϕ es el cuantil $(1-\alpha/2)$ de una curva normal[2].

El sorteo de la muestra estuvo a cargo del Instituto Nacional de Estadística. Las personas relevadas pertenecen a las ciudades de Artigas, Canelones, Ciudad de la Costa, La Paz, Las Piedras, Colonia, Florida, Maldonado, San Carlos, Montevideo, Paysandú, San José, Salto y Tacuarembó. En los casos que no se encontró a la persona se realizó un algoritmo de sustitución de la siguiente manera: "pararse en el punto más noroeste de la manzana y caminar en sentido horario contando el número de casas desde ese punto (casa 1) hasta encontrar una casa con una persona de la edad y sexo requerido" [3](12). La taza de respuesta fue en promedio de 61 %.

[2]Se realizó una muestra probabilística con diseño complejo, el que no será considerado en este enfoque de análisis

[3]Primer Relevamiento Nacional de Salud Bucal en población joven y adulta uruguaya, Lorenzo, S., Álvarez, R., Blanco, S., Peres, M., junio 2013

Apéndice F

Anexo de resultados

Ccorona

Tabla F.1: *Primera Estimación Binomial Negativa*

	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	1.189	0.238	5.001	5.70e-07
Región-Montevideo	-0.056	0.088	-0.643	0.520
Tramo Etario-de 35 a 44	0.123	0.099	1.240	0.215
Tramo Etario-de 65 a 74	-0.794	0.109	-7.301	2.86e-13
Sexo-M	0.098	0.084	1.162	0.245
Estudio Universitario-No	0.232	0.121	1.910	0.056
Institución Médica-No	0.348	0.090	3.882	1.04e-4
Consume Mate-No	-0.318	0.104	-3.041	0.002
Fuma-No	-0.386	0.095	-4.080	4.51e-05
INSE	-0.023	0.004	-5.945	2.77e-09

Tabla F.2: *Primera Estimación Hurdle Binomial Negativa*

Componente Hurdle				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	1.289	0.330	3.910	9.23e-05
Región-Montevideo	-0.100	0.120	-0.837	0.402
Tramo Etario-de 35 a 44	0.093	0.141	0.658	0.510
Tramo Etario-de 65 a 74	-1.104	0.145	-7.632	2.31e-14
Sexo-M	0.113	0.116	0.968	0.333
Estudio Universitario-No	0.130	0.159	0.816	0.414
Institución Médica-No	0.484	0.123	3.940	8.07e-05
Consume Mate-No	-0.379	0.138	-2.750	0.006
Fuma-No	-0.527	0.137	-3.851	1.17e-04
INSE	-0.026	0.005	-4.980	6.41e-07
Componente de Conteo				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	0.919	0.331	2.780	0.005
Región-Montevideo	-0.003	0.115	-0.023	0.982
Tramo Etario-de 35 a 44	0.125	0.123	1.020	0.308
Tramo Etario-de 65 a 74	-0.354	0.164	-2.163	0.030
Sexo-M	0.039	0.112	0.350	0.726
Estudio Universitario-No	0.296	0.173	1.707	0.088
Institución Médica-No	0.217	0.147	-1.637	0.102
Consume Mate-No	-0.241	0.121	-2.390	0.017
Fuma-No	-0.289	0.005	-3.269	0.001
INSE	-0.018	0.225	-2.057	0.040

106

Pcorona

Tabla F.3: *Primera Estimación Cero Inflado Binomial Negativa Pcorona*

Coeficientes del componente Cero Inflado				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	-0.753	0.487	-1.547	0.122
Región-Montevideo	0.373	0.180	2.072	0.038
Tramo Etario-de 35 a 44	-2.251	0.221	-10.18	< 2e-16
Tramo Etario-de 65 a 74	-3.837	0.377	-10.17	< 2e-16
Sexo-M	-0.189	0.181	-1.043	0.297
Estudio Universitario-No	-0.784	0.233	-3.364	0.001
Institución Médica-No	-0.079	0.196	-0.404	0.686
Consume Mate-No	0.643	0.195	3.296	0.001
Fuma-No	0.594	0.212	2.808	0.005
INSE	0.020	0.008	2.561	0.010
Coeficientes del componente de Conteo				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	1.374	0.145	9.465	< 2e-16
Región-Montevideo	-0.037	0.051	-0.741	0.459
Tramo Etario-de 35 a 44	1.516	0.068	22.20	< 2e-16
Tramo Etario-de 65 a 74	2.134	0.072	29.64	< 2e-16
Sexo-M	-0.176	0.050	-3.491	4.82e-04
Estudio Universitario-No	0.002	0.068	0.035	0.972
Institución Médica-No	0.162	0.054	3.002	0.003
Consume Mate-No	-0.091	0.065	-1.417	0.156
Fuma-No	-0.136	0.061	-2.249	0.024
INSE	-0.013	0.002	-5.382	7.36e-08

Tabla F.4: *Primera Estimación Hurdle Binomial Negativa Pcorona*

Coeficientes del componente Hurdle				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	0.464	0.407	1.142	0.254
Región-Montevideo	-0.328	0.152	-2.161	0.031
Tramo Etario-de 35 a 44	2.542	0.195	13.02	< 2e-16
Tramo Etario-de 65 a 74	4.105	0.334	12.27	< 2e-16
Sexo-M	0.066	0.147	0.451	0.652
Estudio Universitario-No	0.673	0.203	3.308	0.001
Institución Médica-No	0.132	0.162	0.814	0.416
Consume Mate-No	-0.599	0.166	-3.613	3.03e-04
Fuma-No	-0.555	0.166	-3.333	0.001
INSE	-0.022	0.006	-3.430	0.001
Coeficientes del componente de Conteo				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	1.372	0.145	9.455	< 2e-16
Región-Montevideo	-0.039	0.051	-0.770	0.441
Tramo Etario-de 35 a 44	1.512	0.068	22.20	< 2e-16
Tramo Etario-de 65 a 74	2.133	0.072	29.69	< 2e-16
Sexo-M	-0.174	0.050	-3.458	5.43e-04
Estudio Universitario-No	0.008	0.068	0.118	0.906
Institución Médica-No	0.167	0.054	3.090	0.002
Consume Mate-No	-0.087	0.064	-1.349	0.177
Fuma-No	-0.136	0.060	-2.255	0.024
INSE	-0.013	0.002	-5.455	4.89e-08

Ocorona

Tabla F.5: *Primera Estimación Cero Inflado Binomial Negativa*

Coeficientes del componente Cero Inflado				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	-0.005	0.437	-0.013	0.989
Región-Montevideo	0.204	0.157	1.297	0.194
Tramo Etario-de 35 a 44	-0.918	0.205	-4.460	8.20e-06
Tramo Etario-de 65 a 74	0.637	0.190	3.346	8.12e-04
Sexo-M	0.284	0.154	1.841	0.065
Estudio Universitario-No	0.646	0.222	2.913	0.003
Institución Médica-No	0.578	0.162	3.548	3.80e-04
Consume Mate-No	-0.220	0.189	-1.168	0.242
Fuma-No	-0.441	0.182	-2.417	0.015
INSE	-0.028	0.007	-3.750	1.71e-04
Coeficientes del componente de Conteo				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	0.599	0.182	3.289	0.001
Región-Montevideo	0.040	0.069	0.578	0.563
Tramo Etario-de 35 a 44	0.759	0.082	9.196	$<$ 2e-16
Tramo Etario-de 65 a 74	0.570	0.093	6.100	1.06e-09
Sexo-M	-0.173	0.069	-2.497	0.012
Estudio Universitario-No	-0.025	0.082	-0.312	0.754
Institución Médica-No	-0.230	0.080	-2.868	0.004
Consume Mate-No	0.039	0.078	0.500	0.616
Fuma-No	0.077	0.086	0.887	0.375
INSE	0.009	0.002	3.617	2.90e-04

Tabla F.6: *Primera Estimación Hurdle Binomial Negativa*

Coeficientes del componente Hurdle				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	-0.442	0.330	-1.339	0.180
Región-Montevideo	-0.166	0.120	-1.381	0.167
Tramo Etario-de 35 a 44	1.157	0.150	7.675	1.65e-14
Tramo Etario-de 65 a 74	-0.237	0.138	-1.714	0.086
Sexo-M	-0.335	0.116	-2.868	0.004
Estudio Universitario-No	-0.519	0.161	-3.228	0.001
Institución Médica-No	-0.574	0.124	-4.622	3.81e-06
Consume Mate-No	0.150	0.137	1.094	0.274
Fuma-No	0.368	0.139	2.655	0.007
INSE	0.024	0.005	4.380	1.19e-05
Coeficientes del componente de Conteo				
	Coeficiente Estimado	Error Estándar	Valor z	P-Valor
(Intercepto)	0.539	0.183	2.938	0.003
Región-Montevideo	0.054	0.069	0.781	0.434
Tramo Etario-de 35 a 44	0.722	0.078	9.197	$<$ 2e-16
Tramo Etario-de 65 a 74	0.536	0.089	6.004	1.93e-09
Sexo-M	-0.154	0.068	-2.249	0.024
Estudio Universitario-No	-0.006	0.082	-0.078	0.937
Institución Médica-No	-0.244	0.079	-3.090	0.002
Consume Mate-No	0.056	0.078	0.722	0.470
Fuma-No	0.092	0.086	1.073	0.283
INSE	0.011	0.002	4.102	4.10e-05